



NOVA

IMS

Information
Management
School

MGI

Mestrado em Gestão de Informação

Master Program in Information Management

Next Best Action nos Seguros

Andreia Braga Duarte Nunes

Trabalho de Projeto apresentada(o) como requisito parcial
para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NEXT BEST ACTION NOS SEGUROS

por

Andreia Braga Duarte Nunes

Trabalho de Projeto apresentado como requisito parcial para a obtenção do grau de Mestre em
Gestão de Informação, Especialização em Gestão de Conhecimento e *Business Intelligence*

Orientador: Professor Doutor Fernando José Ferreira Lucas Bação

Agosto 2020

AGRADECIMENTOS

Gostaria de agradecer aos meus pais pelas oportunidades que me deram ao longo da vida, pelo apoio e por acreditarem sempre em mim.

Aos meus avós, por terem sempre uma palavra de incentivo e por todo o apoio que dão.

Ao meu namorado, Francisco, por me lembrar todos os dias que consigo alcançar tudo aquilo a que me proponho e por estar sempre lá para me ajudar.

À Fidelidade, mas em particular ao Francesco Costigliola, pela oportunidade de desenvolver este projeto e pela paciência que teve ao longo de todo este processo. Agradeço também aos meus colegas do departamento *Center for Advanced Analytics* toda a disponibilidade e conhecimento que transmitiram, sem essa ajuda não seria possível realizar com este projeto.

E, por fim, ao meu orientador Prof. Doutor Fernando Bação pela disponibilidade e ajuda demonstrada.

RESUMO

Com um mercado cada vez mais competitivo e inovador, o setor dos seguros tem vindo a reinventar-se na oferta dos seus produtos e serviços. Este trabalho nasce do projeto *Next Best Action* do departamento *Center for Advanced Analytics* (CAA) da Fidelidade que pretende construir e implementar um sistema de recomendação para facilitar a obtenção de uma visão 360º do cliente. A implementação deste sistema irá permitir oferecer o produto e/ou serviço mais adequado a cada tipo de cliente, tendo em conta as suas necessidades. Para construir uma primeira versão deste sistema de recomendação será necessário desenvolver um modelo de propensão à compra de produtos e/ou seguros não obrigatórios recorrendo a tecnologias SAS.

PALAVRAS-CHAVE

Next Best Action; Modelo de propensão; Seguros; Sistemas de Recomendação

ABSTRACT

With a more competitive and innovative market, the insurance sector has been reinventing itself in the offering its products and services. This work was born in the Next Best Action project of Center for Advanced Analytics (CAA) department at Fidelidade, which intends to build and implement a recommendation system to facilitate the obtainance of a 360 degree view of the client. The implementation of this system will allow to offer the most appropriate product and/or service for each type of customer, taking their needs into account. To build a first version of this recommendation system, it will be necessary to develop a model of propensity to purchase non-compulsory products and/or insurances, using SAS technologies.

KEYWORDS

Next Best Action; Propensity Model; Insurance; Recommender Systems

ÍNDICE

1. Introdução	1
2. Revisão da Literatura	3
2.1. Sistemas de Recomendação	3
2.1.1. Tipos de Sistemas de Recomendação	3
2.1.2. Memory-Based	4
2.1.3. Model-Based	5
2.1.4. Tipo de informação utilizada	6
2.1.5. Aplicação e Vantagens dos Sistemas de Recomendação	6
2.2. Modelos de Machine Learning	7
2.2.1. Árvores de Decisão	7
2.2.2. Regressão Logística	8
2.2.3. Rede neuronal	9
2.2.4. Random Forest	10
3. Next Best Action no contexto da fidelidade	11
3.1. Plano da solução	12
3.1.1. Pilares	12
3.1.2. Coeficientes	12
3.1.3. Aplicação dos Sistemas de Recomendação	13
4. Metodologia	14
4.1. Definição de seguros e/ou coberturas não obrigatórios	15
4.1.1. Tabela de mapeamentos de seguros e/ou coberturas não obrigatórios	16
4.2. Modelo de propensão à compra de produtos e/ou seguros não obrigatórios	17
4.2.1. Identificação do universo e da <i>target</i>	17
4.2.2. Preparação e tratamento de dados	20
4.2.3. Pré-Processamento dos dados	22
4.2.2.1. Métodos de seleção de variáveis	24
4.2.4. Modelação	26
4.2.5. Resultados	27
5. Next Best Action – Aplicação	30
6. Conclusões	32
7. Limitações e Recomendações para Trabalhos Futuros	34
8. Bibliografia	35
9. Anexos	39

ÍNDICE DE FIGURAS

Figura 1 – Tipos de sistemas de recomendação (Fenjiro, s.d.)	3
Figura 2 – Abordagem Híbrida 1	4
Figura 3 - Abordagem Híbrida 2	5
Figura 4 - Abordagem Híbrida 3	5
Figura 5 – Abordagem Híbrida 4	5
Figura 6 – Ilustração de uma árvore de decisão (deVille & Padraic, 2013)	8
Figura 7 – Exemplo de uma rede neuronal	9
Figura 8 – Plano da solução NBA.....	12
Figura 9 – Fases da metodologia SEMMA (SAS Institute)	14
Figura 10 – Parte das cobertas da oferta automóvel no <i>site</i> da Fidelidade	16
Figura 11 – Distribuição do rácio de não obrigatoriedade total.....	19
Figura 12 – Recategorização da variável Descrição da família da apólice em vigor mais antiga	21
Figura 13 - Gráfico de frequência da variável <i>Flag Adulto</i>	21
Figura 14 - Gráfico de frequência da variável “Quantidade de meses para expirar a próxima apólice Automóvel”	22
Figura 15 – Matriz de correlação de <i>Pearson</i> filtrada.....	23
Figura 16 – Diagrama do modelo em SAS Miner	26
Figura 17 - Curva ROC: Treino e Validação.....	27
Figura 18 – Importância das variáveis no modelo final	28
Figura 19 – Valores do impacto das variáveis “Quantidade de apólices em vigor há menos de 2 anos” e “Descrição da NUT 2”	29
Figura 20 – 1º caso do cálculo do <i>score</i> final do NBA	30
Figura 21 - 2º caso do cálculo do <i>score</i> final do NBA.....	31

ÍNDICE DE TABELAS

Tabela 1 - Exemplo da tabela de mapeamento da oferta Fidelidade <i>Pets</i>	17
Tabela 2 – Taxa de propensão do universo no mês de análise.....	19
Tabela 3 – Taxa de propensão do universo final.....	20
Tabela 4 – Nº variáveis por intervalo de percentagem de valores omissos	20
Tabela 5 – Particionamento da tabela	23
Tabela 6 – Categorias das variáveis <i>input</i>	23
Tabela 7 – Estatísticas dos quatro melhores modelos – Treino e Validação	26
Tabela 8 – Categorias das variáveis selecionadas pelo modelo final.....	27

LISTA DE SIGLAS E ABREVIATURAS

ASF	Autoridade de Supervisão de Seguros e Fundos de Pensões
CAA	<i>Center for Advanced Analytics</i>
CF	<i>Collaborative Filtering</i>
CB	<i>Content-based</i>
LOB	<i>Line of Business</i>
ML	<i>Machine Learning</i>
NBA	<i>Next Best Action</i>
PCA	Prémio Comercial Anualizado

1. INTRODUÇÃO

O setor dos seguros é muitas vezes visto como tradicional, mas na verdade é um dos mais inovadores. Dada a natureza da sua atividade, que é a proteção de riscos, tem de antecipar tendências e procurar soluções antes mesmo de elas serem requeridas pelos clientes (Jornal Económico, s.d.).

Hoje em dia, os consumidores distraídos, bombardeados com informação e opções, lutam muitas vezes para encontrar os produtos e serviços que melhor satisfazem as suas necessidades (Davenport, DalleMule, & Lucker, 2011). Como tal, um dos principais objetivos da companhia é o cliente, ou seja, o cliente terá que estar no centro das atenções de tudo o que a companhia faz. É para isso indispensável ter uma visão 360º do mesmo. É necessário conhecê-lo, saber os seus gostos, as suas características, as suas necessidades e o contexto em que se insere. Apenas conhecendo a jornada que cada cliente faz na companhia, é que se consegue melhorar todo o processo de interações e ofertas para que sejam cada vez mais direcionadas e personalizadas.

Com um mercado cada vez mais competitivo e inovador, não bastam apenas as correlações de compra básicas de “As pessoas que compraram isto também compraram aquilo”. Essas correlações não dependem do conhecimento substancial dos atributos do cliente ou do produto e, por isso, são instrumentos pouco precisos. Ofertas um pouco mais direcionadas são baseadas nas compras anteriores de um cliente mas mesmo essas são famosamente indiscriminadas porque, por um exemplo simples, se uma pessoa comprar um CD ou um livro para outra que não partilha dos mesmos gostos que ela, esta ação pode distorcer facilmente as ofertas futuras que essa pessoa irá receber (Davenport, DalleMule, & Lucker, 2011). É necessário encontrar uma solução que ultrapasse essas limitações, uma solução que recolha não só a informação sobre as compras efetuadas, mas também que registre todas as interações que o cliente faça com a companhia ou vice-versa. Só com a recolha de toda a informação disponível sobre o cliente é que se consegue construir uma solução capaz de recomendar a melhor oferta e/ou ação que atenda àquilo que o cliente procure. Fazendo o negócio certo, à hora certa (Davenport, DalleMule, & Lucker, 2011).

O presente trabalho nasce do projeto *Next Best Action* do departamento *Center for Advanced Analytics* (CAA) da Fidelidade no qual estou integrada. Este projeto visa desenhar, planejar e criar as estruturas necessárias que deem suporte ao sistema de recomendação que irá permitir obter uma visão global e transversal de cada cliente, para que se consiga oferecer os produtos e/ou serviços mais adequados. O NBA para além de inovar os processos dos canais de venda, inova também na forma como se interage com o cliente, seguindo um dos motes da Fidelidade: “A inovar para dedicar mais tempo às pessoas, para que a vida não pare”.

O plano deste projeto passa, essencialmente, por 6 etapas:

1. Planeamento e desenho da estrutura do *Next Best Action*;
2. Análise e recolha da informação dos modelos desenvolvidos que podem incorporar o NBA;
3. Análise e recolha da oferta existente, e a respetiva catalogação dos seguros e/ou coberturas como obrigatórias ou não obrigatórias;
4. Desenvolvimento do modelo de propensão à compra de produtos e/ou coberturas não obrigatórias, que retorna o nível de sensibilidade que cada cliente tem em relação ao tema dos seguros;

5. Criação da 1ª versão do sistema de recomendação;
6. Definição dos próximos passos para a melhoria e robustez do NBA.

A nível técnico recorreu-se a dois dos *softwares* do *SAS Institute*: *SAS Enterprise Guide*, para a criação das tabelas de mapeamentos que definem quais os seguros e/ou coberturas obrigatórias ou não obrigatórias. Tabelas essas que alimentarão o universo necessário para o modelo de propensão à compra de produtos e/ou seguros não obrigatórios bem como para a construção da estrutura do sistema de recomendações. Por fim, o *SAS Enterprise Miner* para toda a parte de modelação referente ao modelo de propensão.

2. REVISÃO DA LITERATURA

2.1. SISTEMAS DE RECOMENDAÇÃO

Os sistemas de recomendação (SR) são ferramentas de *software* e técnicas que podem ser usadas para fornecer sugestões de itens. Estas sugestões estão relacionadas com diversos processos de tomada de decisão, tais como, que itens comprar, que música ouvir ou que notícias *online* ler. “Item” é um termo geral usado para indicar o que é que o sistema recomenda ao consumidor (Mitra, Chaudhari, & Patwardhan, 2014).

As recomendações podem assumir múltiplas formas. Recomendações não personalizadas que é o caso da sugestão de livros, revistas, CDs, entre outros, ou recomendações personalizadas que são possíveis através do *e-commerce* ou dos *websites*, onde o utilizador tem sugestões específicas através da análise das suas interações prévias com o *website* e, se disponível, o *feedback* dos seus círculos sociais (Mitra, Chaudhari, & Patwardhan, 2014). Na sua forma mais simplista as recomendações personalizadas podem ser oferecidas como listas ordenadas de itens. Para realizar esta ordenação, os SR tentam prever qual o produto ou serviço mais indicado para cada cliente, baseado nas suas preferências e restrições. Estes sistemas geram as suas recomendações usando vários tipos de conhecimento e dados sobre os consumidores, os itens disponíveis e as compras/transações anteriormente realizadas (Ricci, Rokach, Shapira, & Kantor).

2.1.1. TIPOS DE SISTEMAS DE RECOMENDAÇÃO

Para escolher o sistema mais adequado ao NBA, é necessário conhecer que tipos de algoritmos de SR existem. Na figura 1 está representado um esquema que resume bem todos os algoritmos existentes:

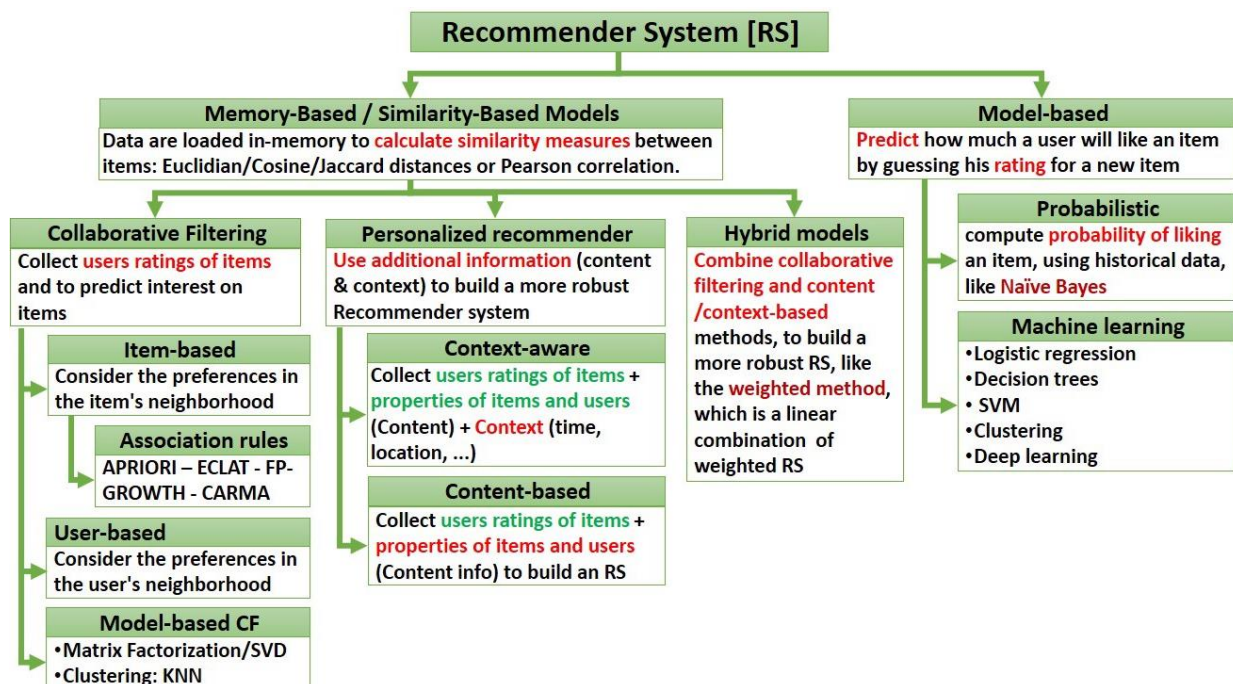


Figura 1 – Tipos de sistemas de recomendação (Fenjiro, s.d.)

2.1.2. MEMORY-BASED

Os modelos *memory-based* são modelos onde se calcula a similaridade entre utilizadores e itens usando técnicas de vizinhança (métricas de similaridade). As mais comuns são a distância Euclidiana, a similaridade *Cosine* e o coeficiente de correlação de *Pearson* (Fenjiro, s.d.). Dentro destes modelos temos o *Collaborative Filtering (CF)*, *Content-based (CB)* e os modelos híbridos.

A maior parte dos SR desenvolvidos são baseados nas técnicas de CF que identificam as recomendações por utilizador, comparando o seu perfil, que é um modelo das preferências e interesses do utilizador, com os perfis de outros utilizadores, propondo de seguida os itens de perfis semelhantes (Gurbanov & Ricci, 2017). As vantagens desta abordagem é a implementação dos SR se tornar mais fácil, o facto de se poder adicionar novos dados de forma simples e incremental e a melhoria na *performance* da previsão. Por outro lado, estes sistemas requerem muitas vezes uma enorme quantidade de dados, onde o utilizador pode fazer recomendações exatas, necessitam de uma enorme capacidade computacional para calcular as recomendações, pois são calculadas para diversos ambientes onde existem biliões de utilizadores e produtos e, por fim, o número de itens vendidos é enorme mas apenas um pequeno conjunto é classificado pela maioria dos utilizadores ativos, logo são poucas as classificações dadas aos itens mais populares (Thorat, Goudar, & Barve, 2015).

Outra das técnicas é o *Content-based (CB)* onde os SR aprendem a recomendar os itens que são semelhantes aos que o utilizador classificou positivamente no passado. A semelhança dos itens é calculada através das características associadas com os itens comparados (Ricci, Rokach, Shapira, & Kantor). A descrição e o perfil da orientação dos utilizadores desempenham um papel muito importante no CB (Thorat, Goudar, & Barve, 2015). Como vantagens, fornece independência através dos *ratings* exclusivos que são usados pelo utilizador ativo para construir o seu próprio perfil, fornece transparência aos utilizadores ativos explicando como é que o SR funcionam, e é uma técnica adequada para recomendar itens que não foram recomendados a nenhum utilizador, o que é vantajoso para novos consumidores. Já como limitações, aparenta ser uma tarefa difícil para gerar atributos para itens em certas áreas, advoga os mesmos tipos de itens, pois sofre de um problema de superespecialização e, por fim, é difícil de adquirir *feedback* dos utilizadores porque os mesmos não costumam classificar os itens, como no CF, logo não é possível determinar se a classificação dada é correta (Thorat, Goudar, & Barve, 2015).

Por último, temos os modelos híbridos onde os SR são baseados na combinação das técnicas acima mencionadas. Um sistema híbrido combina as técnicas do CF e do CB e tenta usar as vantagens de um para corrigir as limitações do outro. A abordagem híbrida é utilizada para melhorar a precisão das recomendações e pode ser implementada de diversas formas (Thorat, Goudar, & Barve, 2015). A figura 2 mostra a implementação das técnicas CF e CB individualmente, combinando posteriormente as suas previsões para produzir melhores recomendações (Thorat, Goudar, & Barve, 2015).

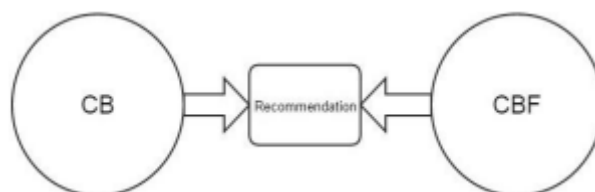


Figura 2 – Abordagem Híbrida 1

Na figura 3, temos a 2ª abordagem híbrida que integra algumas características de CB dentro da abordagem CF. Esta abordagem ultrapassa a limitação do CF que requer uma enorme quantidade de dados para produzir recomendações exatas e o problema de superespecialização do CB (Thorat, Goudar, & Barve, 2015).

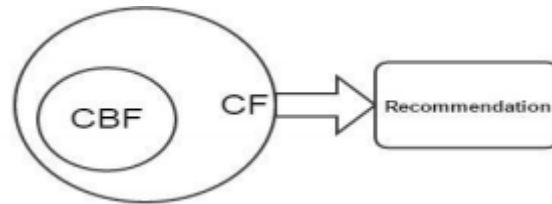


Figura 3 - Abordagem Híbrida 2

A abordagem híbrida 3, na figura 4, ilustra a incorporação de algumas características CF dentro da abordagem CB (Thorat, Goudar, & Barve, 2015).

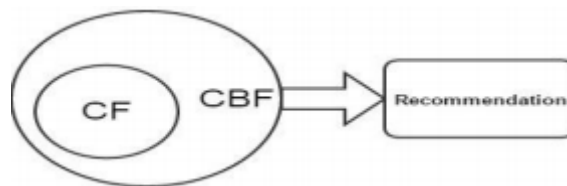


Figura 4 - Abordagem Híbrida 3

E a figura 5, mostra um modelo construído através da unificação de características CF e CBF. Este modelo unificado melhora a eficácia do processo de recomendação (Thorat, Goudar, & Barve, 2015).

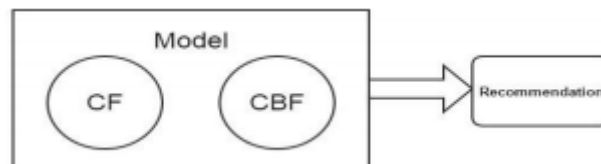


Figura 5 – Abordagem Híbrida 4

Como já foi referido anteriormente, a necessidade de uma grande quantidade de dados e a escassez de itens classificados, são dois problemas bem comuns nos SR que podem ser ultrapassados utilizando uma das quatro abordagens acima apresentadas. Um ótimo exemplo de um SR Híbrido é a Netflix, pois esta plataforma de *streaming* apresenta recomendações baseadas na exploração dos hábitos de utilizadores similares (CF) mas também fornece filmes com características que são partilhadas com os filmes em que os utilizadores atribuem classificações mais altas (CB) (Thorat, Goudar, & Barve, 2015).

2.1.3. MODEL-BASED

As técnicas *model-based* utilizam a fase do treino do modelo para aprender os parâmetros e padrões utilizando um algoritmo de otimização como o *gradient descent*, sem necessitar de utilizar todo o *dataset*. Em vez de calcular a matriz para estimar as preferências dos utilizadores, estas técnicas utilizam a pesquisa do vizinho mais próximo para obter uma matriz aproximada, utilizando algoritmos de *machine learning* (ML), tais como, as redes neuronais como funções aproximadoras. O uso das

técnicas *model-based* traz benefícios na velocidade de processamento e na escassez de dados (Fenjiro, s.d.).

2.1.4. TIPO DE INFORMAÇÃO UTILIZADA

O *input* para os SR pode variar conforme o tipo de algoritmo e a área de atuação escolhida mas geralmente os dados de *input* distribuem-se pelo comportamento do cliente e/ou *prospect*, onde estão incluídas as transações efetuadas, a utilização do cliente e os interesses do mesmo. Como exemplo disso são as compras e subscrições realizadas numa operadora de telecomunicações (Fenjiro, s.d.). Também os detalhes dos itens servem como *input*, ou seja, características e atributos dos mesmos, por exemplo: tipo de produto, preço do mesmo, género e idade do cliente, entre outras. Para além disto, poderá estar incluído os *Ratings* dados pelo cliente, a disponibilidade e informação do inventário (Fenjiro, s.d.). Por último, temos a informação contextual onde poderá estar incluída a informação sobre o tempo (mês, ano, etc.), a localização (país, região, clima, linguagem, etc.), tipo de pessoa (se é relutante, impulsivo, recetiva, etc.) e toda a informação sobre as redes sociais (número de gostos, partilhas, recomendações, etc.) (Fenjiro, s.d.).

2.1.5. APLICAÇÃO E VANTAGENS DOS SISTEMAS DE RECOMENDAÇÃO

Os SR têm um papel bastante importante no mundo digital, a Netflix e a Amazon são dois dos muitos exemplos que utilizam este tipo de sistemas para melhorar o serviço fornecido aos seus clientes, ou seja, as sugestões personalizadas para cada cliente baseadas no seu comportamento. Para além do mundo digital, os SR têm vindo a ser adotados pelo mercado retalhista, financeiro ou pela área da saúde, sendo que começa a dar os seus primeiros passos no mercado segurador. Isto porque, tal como foi dito anteriormente, o setor dos seguros é um setor bastante tradicional mas que tem tentado inverter essa tendência, procurando novas soluções para se inovar e para servir o cliente da melhor maneira possível. Uma das razões que pode levar à aplicação dos SR no setor segurador é o aumento da receita. Este deve ser o aspeto comercial mais importante dos SR, pois as empresas conseguem vender mais produtos e/ou serviços do que quando não tinham qualquer tipo de recomendação. Este objetivo é atingido, pois os itens recomendados são adequados às necessidades dos clientes (Ricci, Rokach, Shapira, & Kantor). Outra das razões é o aumento da satisfação do cliente pois, com um SR bem desenhado, os clientes obtêm recomendações que acham interessantes e relevantes para as suas necessidades. Logo a eficiência das recomendações aumenta bem como a probabilidade de aceitação das mesmas, o que leva à satisfação dos clientes (Ricci, Rokach, Shapira, & Kantor). O aumento da fidelidade do cliente é outro aspeto a ter em conta pois, um consumidor deve ser leal à companhia, se cada vez que haja uma interação, a empresa reconhecer que é um cliente antigo. Como os SR utilizam informações de interações anteriores para realizar uma recomendação, quanto mais vezes o cliente interagir com a empresa, mais informação é adquirida e mais refinado o sistema se torna, gerando recomendações cada vez mais personalizadas e adequadas às preferências do consumidor (Ricci, Rokach, Shapira, & Kantor). De seguida, a penúltima razão recai sobre perceber o que o cliente realmente quer, ou seja, com o histórico de transações e outros tipos de dados sobre o cliente, o sistema detém o conhecimento para recomendar produtos e/ou serviços que satisfazem as necessidades do cliente. Em alguns casos, o SR pode ter dificuldades em fazer boas recomendações quando existe um número relativamente pequeno de itens ou quando o serviço é de uma área crítica como a da saúde ou financeira (Ricci, Rokach, Shapira, & Kantor). E, por último, vender mais itens diversificados. Com base nos gostos de cada cliente, podem ser recomendados itens que este normalmente não compraria mas que são complementares ou que vão de encontro às necessidades

do cliente, um exemplo disso são as recomendações que a Netflix faz. Para além das sugestões dos filmes e séries mais populares, a Netflix sugere outros itens menos conhecidos ou populares com base nos gostos do utilizador, sem estas sugestões o consumidor poderia nunca vir a conhecer os mesmos (Ricci, Rokach, Shapira, & Kantor).

2.2. MODELOS DE MACHINE LEARNING

Com vista à previsão da propensão dos clientes a adquirir determinado produto, serão utilizados diversos algoritmos de *machine learning*. ML é um termo genérico que se refere a um vasto leque de algoritmos que realizam previsões inteligentes baseadas em *datasets* normalmente grandes (Nichols, Chan, & Baker, 2018). Os métodos de ML são geralmente flexíveis e não-paramétricos, permitindo efetuar previsões ou classificações de dados (D., Kirchner, Eck, & Signorino, 2018). Normalmente, as técnicas de ML podem ser divididas em duas amplas categorias, supervisionadas ou não supervisionadas. A aprendizagem supervisionada, como o nome sugere, envolve ensinar o modelo com uma coleção de dados *input* que têm o *output* correto já associado (Nichols, Chan, & Baker, 2018). Existem dois tipos de aprendizagem supervisionada, a primeira, é a classificação, onde os dados são distribuídos em categorias definidas no *dataset* de acordo com as suas características específicas (ÇELİK & ALTUNAYDIN, 2018). O segundo tipo, a regressão, é a previsão de uma variável contínua (Nichols, Chan, & Baker, 2018). A aprendizagem não supervisionada, por outro lado, é mais complexa, uma vez que não existe uma variável dependente pré-definida e os seus métodos focam-se na deteção de padrões ao longo de todas as variáveis de interesse no *dataset* (D., Kirchner, Eck, & Signorino, 2018). Neste projeto, o modelo de propensão que irá ser desenvolvido é um modelo de classificação baseado na aprendizagem supervisionada, visto que, o *dataset* irá ter uma variável dependente binária pré-definida, e os algoritmos de ML utilizados serão árvores de decisão, regressão logística, rede neuronal e *random forest*.

2.2.1. ÁRVORES DE DECISÃO

As árvores de decisão são uma das mais poderosas técnicas de *data mining* porque se pode usar num enorme leque de problemas e porque produzem modelos que podem explicar como é que elas funcionam (Linoff & Berry, Chapter 7 - Decision Trees, 2011). As árvores de decisão são produzidas por algoritmos que identificam diferentes maneiras de dividir o *dataset* em segmentos tipo ramos. Esses segmentos formam uma árvore de decisão invertida que origina o nó raiz no topo da árvore. O objeto de análise é refletido neste nó raiz como uma exibição simples e unidimensional na interface da árvore de decisão. A figura 6 mostra um exemplo de uma árvore de decisão onde a descoberta de uma regra de decisão para formar ramos ou segmentos sob o nó raiz, é baseada num método que extrai a relação entre o objeto de análise (que serve como a variável *target* nos dados) e uma ou mais variáveis, que servem como variáveis *input*, para criar os ramos ou segmentos. Os valores nos campos *input* são usados para estimar o valor provável da *target*. Uma vez extraída a relação, uma ou mais regras de decisão que descrevem as relações entre *inputs* e *targets* podem ser obtidas. As regras podem ser selecionadas e utilizadas para exibir a árvore de decisão, que fornece um meio de examinar e descrever visualmente a árvore como uma rede de relações que caracteriza os valores das variáveis *input* e a *target*. As regras de decisão podem prever os valores de observações novas ou nunca vistas que contêm valores para os *inputs*, mas podem não conter valores para a *target* (deVille & Padraic, 2013).

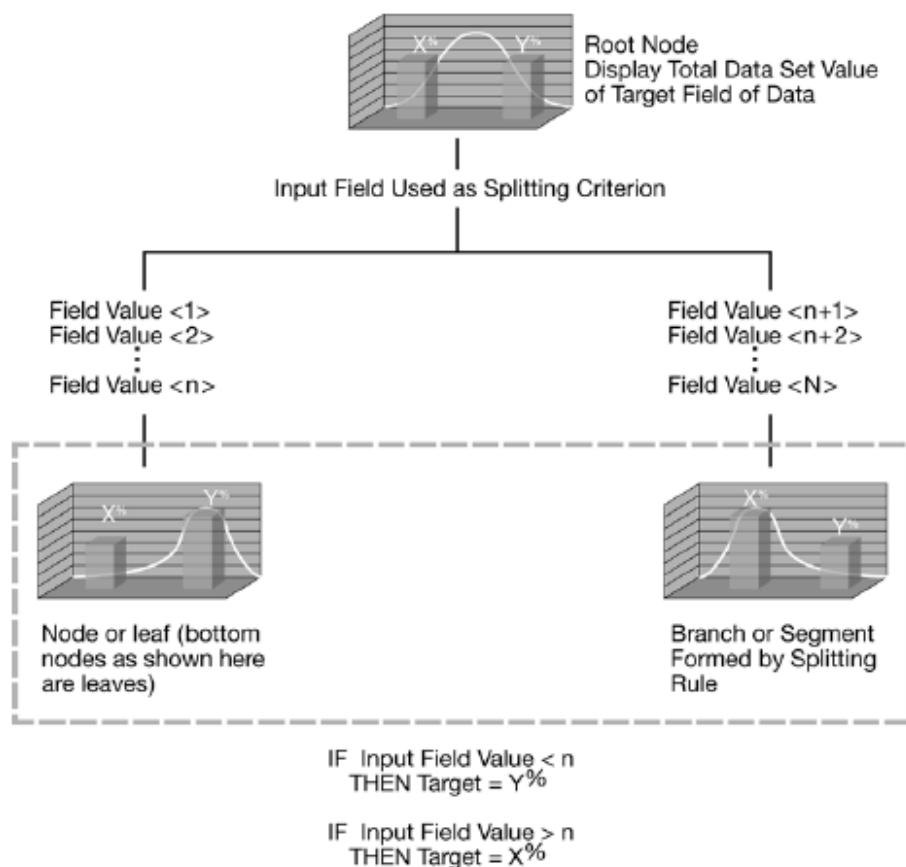


Figura 6 – Ilustração de uma árvore de decisão (deVille & Padraic, 2013)

Resumidamente, uma árvore de decisão é uma coleção hierárquica de regras que descrevem como dividir um grande grupo de registos em grupos sucessivamente menores. Com cada divisão sucessiva, os membros dos segmentos resultantes tornam-se cada vez mais semelhantes entre si, em relação à *target* (Linoff & Berry, Chapter 7 - Decision Trees, 2011).

2.2.2. REGRESSÃO LOGÍSTICA

Em termos mais simples, a regressão é um procedimento estatístico que tenta prever os valores de uma dada variável, denominada de dependente, resultado ou variável de resposta, com base nos valores de uma ou mais variáveis chamadas de independentes, preditores ou co-variáveis. O resultado de uma regressão é geralmente uma equação (ou modelo) que sumariza a relação entre a variável dependente e as independentes (Guido, Winters, & Rains, 2006). Os métodos de regressão logística são aplicados a qualquer estudo onde o resultado de interesse seja dicotómico, isto é, ou o resultado ocorre ou não ocorre. Na maioria das vezes, ao descrever se o resultado ocorre, a notação binária é utilizada para denotar a sua presença (resultado=1) ou ausência (resultado=0) (Rasouliyan & Miller). Uma vez que o resultado é dicotómico, a previsão de alterações na unidade tem pouco ou nenhum significado. Como alternativa para modelar o valor do resultado, a regressão logística concentra-se na probabilidade relativa (*odds*) de obter determinada categoria de resultados (Guido, Winters, & Rains, 2006).

2.2.3. REDE NEURONAL

As redes neurais receberam esse nome devido à sua similaridade com a forma como os neurónios funcionam no corpo humano (Lavery & Mawr, 2016). Uma rede neuronal é um conjunto de variáveis *input/output* conectadas, onde cada conexão tem um peso associado, são uma classe de modelos paramétricos que podem capturar uma variedade de relações não lineares entre um conjunto de preditores e a variável *target*. Construir uma rede neuronal envolve dois passos principais, em primeiro lugar, deve-se definir a configuração ou estrutura da rede, e depois treinar iterativamente o modelo baseado nessa dada estrutura (Zhao).

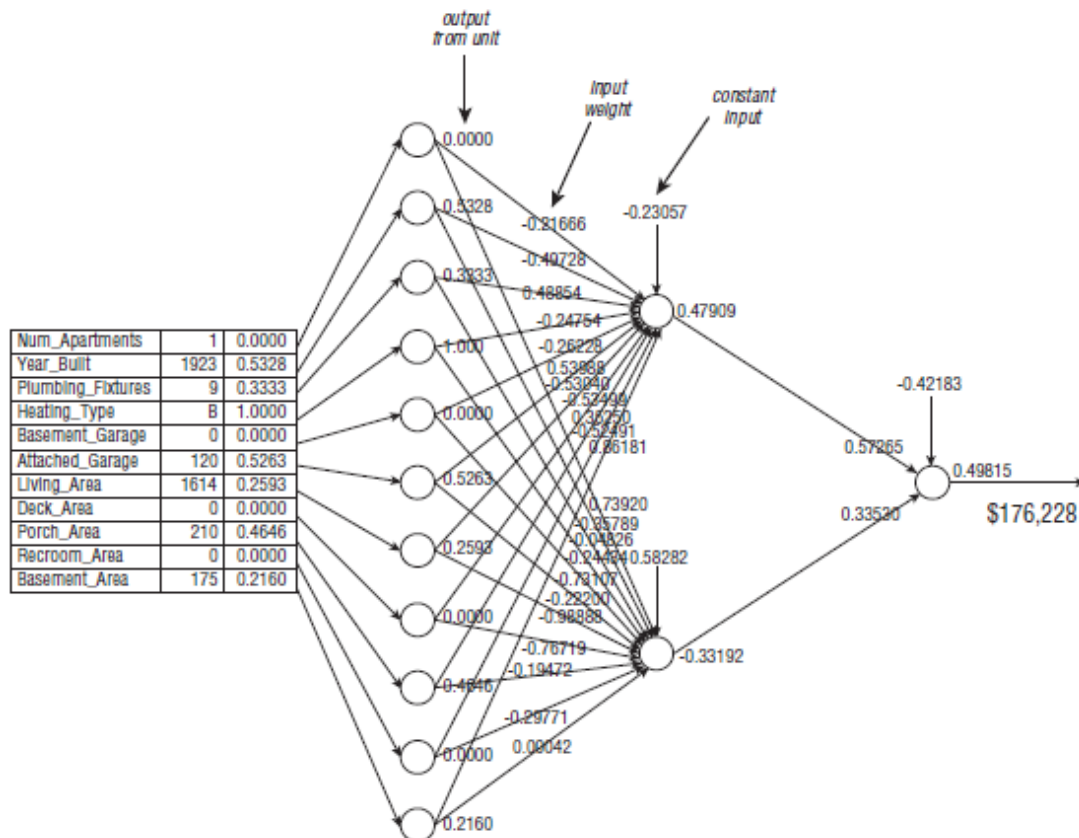


Figura 7 – Exemplo de uma rede neuronal

A figura 7 ilustra um exemplo de rede neuronal que representa um modelo para estimar valores imobiliários reais. A topologia, ou estrutura, desta rede é típica nas redes utilizadas para previsão e classificação. As unidades são organizadas em três camadas. A camada da esquerda está conectada com os *inputs* e chama-se camada *input*. A camada oculta não está conectada nem aos *inputs* nem aos *outputs* da rede. Cada unidade da camada oculta é completamente conectada a todas as unidades da camada *input*. Como esta rede contém unidades *standard*, as unidades da camada oculta calculam o seu *output* através da multiplicação do valor de cada *input* com o respetivo peso, somando-os e aplicando a função de transferência. Uma rede neuronal pode ter um qualquer número de camadas ocultas mas em geral, uma camada é suficiente. Quanto mais larga for a camada, ou seja, quanto mais unidades contiver, maior a capacidade da rede em reconhecer padrões. Na figura 7 também é possível visualizar um *input* adicional vindo do topo da rede, este *input* denomina-se de *input* constante, ou *bias*, e é sempre definido como 1. O *bias* atua como uma compensação global que

ajuda a rede a capturar melhor os padrões. A última unidade da direita é a camada de *output* porque está conectada com o *output* da rede neuronal e, também está totalmente conectada a todas as unidades da camada oculta (Linooff & Berry, Chapter 8 - Artificial Neural Networks, 2011).

2.2.4. RANDOM FOREST

O modelo *Random Forest* é um modelo preditivo que consiste em diversas árvores de decisão que diferem entre si em duas maneiras. Em primeiro lugar, os dados de treino para uma árvore são uma amostra sem reposição de todas as observações disponíveis. Em segundo lugar, as variáveis *input* que são consideradas para a divisão do nó, são selecionadas aleatoriamente de todas as variáveis *input* disponíveis. Os dados de treino para uma árvore individual excluem alguns dos dados disponíveis. Os dados que são retidos durante o treino são chamados de amostra *out-of-bag*. Uma árvore individual apenas usa a amostra *out-of-bag* para formar previsões, visto que são mais confiáveis do que as previsões dos dados de treino (SAS, Random Forest Model, s.d.).

A correlação entre árvores é reduzida ao selecionar aleatoriamente as variáveis, o que melhora o poder de previsão, resultando numa maior eficiência. As vantagens da aplicação das *random forests*, são a superação do problema de *overfitting*, o facto de nos dados de treino serem menos sensíveis aos *outliers*, os parâmetros poderem ser facilmente definidos, eliminando a necessidade de podar as árvores e, por fim, a geração automática da importância das variáveis e da precisão (Ali, Khan, Ahmad, & Maqsood, 2012).

3. NEXT BEST ACTION NO CONTEXTO DA FIDELIDADE

Em primeiro lugar, é necessário perceber qual o processo implementado anteriormente e que fez com que fosse necessário criar uma solução alternativa.

O aumento das vendas, a diversificação dos produtos vendidos e o fortalecimento das relações com os clientes, são três dos muitos objetivos de qualquer empresa, neste caso, de uma seguradora. Para atingir esses objetivos é essencial conhecer os seus clientes, as necessidades dos mesmos e perceber qual o melhor momento para os abordar com uma nova oferta. Para isso são aplicadas algumas estratégias de *marketing* baseadas em modelos de *cross-sell*, *up-sell* e a retenção de clientes, onde o processo anterior se assentava.

O *Cross-sell* define-se como uma venda adicional de produtos que estão relacionados com outros comprados anteriormente, enquanto o *up-sell* envolve o aumento do volume da compra através da venda de mais unidades do mesmo produto ou fazendo o *upgrade* para uma versão mais dispendiosa da que foi comprada (Kamakura, 2007). Numa seguradora, um dos exemplos de *cross-sell* é a oferta de um seguro de vida risco para quem já tenha um seguro de multirriscos, isto porque, a maioria das pessoas que compra uma casa necessita de fazer um empréstimo bancário que leva à compra desses dois tipos de seguro, mas nem sempre os seguros são comprados na mesma seguradora, por isso, esta é uma combinação de *cross-sell* bastante frequente. No caso do *up-sell*, um dos exemplos é o *upgrade* de coberturas nos seguros automóveis. Para as pessoas que compram o plano base deste seguro, é-lhes proposto o *upgrade* para o plano seguinte, no caso da Fidelidade, seria a passagem do plano Auto 1 para o Auto 2 que já inclui a cobertura de Quebra Isolada de Vidros.

O nível de *Cross-sell* é o mais completo até ao momento, visto que contém modelos desenvolvidos para as LOBS¹ de Auto, Habitação, Saúde, Poupança, Viagem e Lazer (Acidentes Pessoais) e Família (Vida Risco). Para completar este nível apenas falta o modelo de *Pets*. Todas as listas de potenciais compradores geradas por estes modelos, ou por outras palavras, *leads*, são distribuídas posteriormente para os mediadores da Fidelidade. Quanto às ofertas de *Up-Sell* não são sempre as mesmas, variam conforme as iniciativas da companhia. Porém, quando as ofertas correspondem aos modelos de *cross-sell* acima mencionados, são utilizadas as *leads* geradas pelos mesmos. Caso não exista um modelo correspondente à oferta, pode-se recorrer a duas abordagens, na primeira, é criado um modelo específico para a oferta e, na segunda, caso não exista informação suficiente que suporte o desenvolvimento de um modelo, são aplicados filtros que exemplificam as regras de negócio mais apropriadas.

Por fim, quando um cliente para de fazer negócios com uma empresa, esse cliente está a fazer *churn* e, para evitar que isso aconteça, são realizadas atividades de retenção de clientes. O objetivo destas atividades é ajudar a que a companhia retenha o máximo de clientes possível (Galletto, 2015). Na realidade da Fidelidade, a retenção de clientes pode passar pela oferta de um desconto no prémio do seguro ou, em relembrar o cliente de que tem uma consulta de *check-up* gratuita, no caso dos seguros de saúde. O nível de *Churn* implementado no processo anterior apenas apresenta o modelo desenvolvido para a LOB Saúde, em que os resultados alimentam a iniciativa de retenção de clientes realizada pela Multicare.

¹ Os nomes das LOBS seguem a apresentação do site www.fidelidade.pt

3.1. PLANO DA SOLUÇÃO



Figura 8 – Plano da solução NBA

O sistema de recomendação *Next Best Action* (NBA), ilustrado pela figura 8, foi planeado para usar a figura completa do cliente particular, em termos de perfil, necessidades e comportamentos atuais. O objetivo final é ter uma solução de *machine learning* que seja multicanal e personalizada. Ou seja, em vez de estarmos a importunar o cliente com múltiplas ofertas e/ou ações, o NBA recomenda automaticamente qual a melhor ação e/ou oferta a realizar com o mesmo.

3.1.1. Pilares

Este sistema assenta em três pilares base: Retenção, Produtos & Serviços e Ações. Estes pilares são encarados como “independentes” em termos de desenvolvimento. O primeiro pilar, Retenção, foca-se em fornecer modelos preditivos de retenção de clientes de forma a gerar iniciativas preventivas. Para tornar este pilar robusto é necessário desenvolver modelos para todas as linhas de negócio de forma a ter uma visão 360º do cliente. No pilar Produtos & Serviços estão representados um leque de potenciais ofertas para os clientes ativos da companhia. Estas ofertas são baseadas nos modelos de *cross-sell* e *up-sell* já disponíveis. Por último, o pilar Ações, é referente a todas as potenciais ações direta ou indiretamente ligadas com os produtos e serviços existentes. Estas ações são variadas e podem ser inquéritos de satisfação do cliente ou simples chamadas informativas.

3.1.2. Coeficientes

Com vista a completar o NBA foram adicionados três coeficientes que são globais a todo o modelo. Em primeiro lugar, temos o ciclo comercial onde é definido quais os produtos estratégicos para a companhia, de forma a encaminhar a ação comercial dos mediadores. No processo

implementado anteriormente apenas eram geradas *leads* para as LOBS que estavam em ciclo, mas como o NBA irá englobar todas as LOBS existentes, é preciso criar um ponderador que dê um peso maior aos produtos que estão em ciclo naquele instante. Assim, o coeficiente do ciclo comercial é parametrizável e pode conter valores entre 0 e 1. Neste momento, o peso atribuído para os produtos em ciclo é de 0,80 e para os restantes é de 0,20. O coeficiente da especificação de produtos tem como objetivo ajudar a definir qual a oferta que mais se adequa a cada cliente. Para este coeficiente foi construído um modelo de propensão à compra de seguros e/ou produtos não obrigatórios que transmitem a sensibilidade geral do cliente à temática dos seguros, ou seja, se são clientes mais sensíveis ao tema, irão prestar mais atenção às coberturas e não tanto ao seguro mais básico que apenas cobre as coberturas obrigatórias por lei. O desenvolvimento deste modelo irá ser especificado de seguida. Por fim, o coeficiente de retenção é o que determina o quão importante é fazer um contacto preventivo em vez de oferecer um novo produto e/ou serviço, e é baseado no valor que o cliente tem para a companhia. Por exemplo, se for um cliente com um seguro de Multirriscos Habitação com cobertura de sismo, que é uma cobertura não obrigatória e que encarece bastante o PCA, e se o mesmo tiver uma elevada propensão em fazer *churn*, será que vale a pena oferecer um novo produto de outra LOB, ou tentar retê-lo? Este coeficiente irá ajudar nesta decisão.

3.1.3. Aplicação dos Sistemas de Recomendação

Para além dos pilares e coeficientes acima descritos, é importante referir qual o tipo de sistema de recomendação a ser utilizado. Para este caso específico, o ideal será aplicar um modelo híbrido *memory-based* aplicando a abordagem 4, acima representada pela figura 5, onde são combinadas as técnicas de CF e CB, para melhorar a precisão da recomendação personalizada de produtos, serviços ou ações para cada cliente. Na prática, o pilar dos Produtos & Serviços incorpora as técnicas de CF através dos modelos de *up-sell*, pois irá analisar as preferências passadas do utilizador e comparar com outros que tenham perfil semelhante para propor novas coberturas. Para os modelos de *cross-sell*, irão ser aplicadas as técnicas CB onde o conhecimento sobre um determinado produto é fundamental para recomendar novos produtos baseados nos padrões de consumo anteriores. No pilar da Retenção também serão aplicadas técnicas CB, uma vez que para perceber quais os clientes que vão fazer *churn*, compara-se perfis de utilizadores que abandonaram no passado e procura-se perfis semelhantes de clientes para propor. Por fim, no pilar Ações, as técnicas a ser aplicadas são as CF, pois para prever qual a próxima ação a realizar com cada cliente, utiliza-se o conhecimento adquirido anteriormente sobre os padrões de interações de clientes com perfis semelhantes.

O modelo de recomendação híbrido *memory-based* tal como foi mencionado acima, não é possível de aplicar de momento, visto que a empresa não reúne toda a informação necessária para aplicar um sistema deste tipo. Atualmente apenas é exequível aplicar um sistema de recomendação *model-based* utilizando modelos de *machine learning* para prever a propensão de cada cliente à compra de um determinado produto. Desses modelos fazem parte todos os modelos integrantes do pilar Produtos & Serviços. De seguida, tendo todas as propensões, é possível calcular qual a oferta mais adequada para cada cliente.

4. METODOLOGIA

Após estruturar o sistema de recomendação, é necessário analisar qual o caminho inicial mais exequível para começar a criar a estrutura definida. O trabalho que origina este projeto baseia-se na 1ª versão do NBA, no qual o plano de ataque passa pela componente mais completa até então, o pilar dos Produtos & Serviços, onde temos praticamente todos os modelos de *cross-sell* desenvolvidos. Quanto aos coeficientes que são globais a todo o NBA, e depois de analisar toda a informação disponível até ao momento, a estratégia passa por criar os coeficientes do ciclo comercial e o coeficiente da especificação de produtos. Para criar o último coeficiente é necessário, tal como foi referido anteriormente, um modelo de propensão aos produtos e/ou seguros não obrigatórios cuja base de desenvolvimento será através da utilização dos *softwares* SAS Enterprise Guide e SAS Enterprise Miner.

Para a criação do modelo recorreu-se à metodologia SEMMA desenvolvida pelo SAS Institute. Esta metodologia oferece um processo de fácil compreensão, permitindo um desenvolvimento e manutenção organizada e adequada para projetos de *Data Mining*. Confere assim, uma estrutura para a sua conceção, criação e evolução, ajudando a apresentar soluções para os problemas de negócio bem como a encontrar os objetivos de negócio de *Data Mining* (Azevedo & Santos , 2005). SEMMA (Sample, Explore, Modify, Model, and Assess) divide o processo de *Data Mining* em cinco fases: a primeira é a *Sample*, onde existe uma amostragem dos dados criando uma ou mais tabelas. As amostras devem ser grandes o suficiente para conter informação significativa, mas pequenas o suficiente para processar rapidamente. Passando para a segunda fase, *Explore*, que consiste na exploração dos dados através da pesquisa antecipada por relações, tendências imprevistas e anomalias, a fim de obter compreensão e ideias. A terceira fase, *Modify*, modifica os dados através da criação, seleção e transformação das variáveis com foco no processo de seleção do modelo. Segue-se a quarta fase, *Model*, onde se modela os dados permitindo que o *software* pesquise automaticamente por uma combinação de dados que prevejam o resultado esperado. E, por fim, na fase *Assess*, avaliam-se os dados através da avaliação da utilidade e confiabilidade das descobertas do processo de *Data Mining* (SAS Institute).

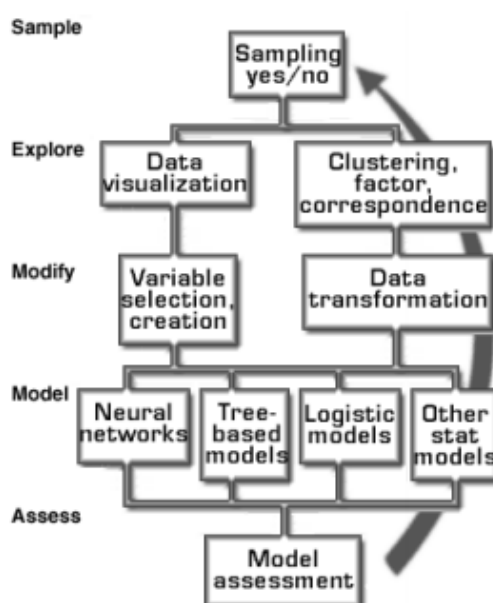


Figura 9 – Fases da metodologia SEMMA (SAS Institute)

4.1. DEFINIÇÃO DE SEGUROS E/OU COBERTURAS NÃO OBRIGATÓRIOS

Para iniciar o modelo de propensão e antes da identificação do Universo e da criação da *Target*, é necessário perceber quais os seguros e/ou coberturas que todas as pessoas são obrigadas a ter. Neste caso, o interesse apenas reside nos seguros obrigatórios para pessoas individuais visto ser a população alvo tanto para o modelo de propensão como para o NBA.

Assim, segundo a ASF, os seguros obrigatórios para particulares recaem sobre o seguro de Animais, onde o detentor de qualquer animal perigoso ou potencialmente perigoso fica obrigado a possuir um seguro de responsabilidade civil destinado a cobrir os danos causados por este (Artigo 10º, Decreto-Lei n.º 315/2009, s.d.), com um capital mínimo de 50 000€. De seguida, temos o seguro automóvel que é, sem qualquer margem para dúvidas, o mais popular de todos os seguros e o mais representado na estrutura das carteiras de qualquer seguradora em Portugal, no que respeita a seguros do ramo não vida (Gilberto, 2012). Toda a pessoa que possa ser civilmente responsável pela reparação de danos corporais ou materiais causados a terceiros por um veículo terrestre a motor para cuja condução seja necessário um título específico e seus reboques, com estacionamento habitual em Portugal, deve, para que esses veículos possam circular, encontrar-se coberta por um seguro que garanta tal responsabilidade, nos termos do presente decreto-lei (Artigo 4º, Decreto-Lei nº 291/2007, s.d.). O capital mínimo obrigatório a partir de 1 de Junho de 2012 é de 5 000 000€ por acidente de danos corporais e de 1 000 000€ por acidente de danos materiais (Artigo 12º, Decreto-Lei n.º 291/2007, s.d.). A posse de uma arma de fogo é sempre passível de provocar acidentes mais ou menos graves e a caça é uma das atividades propícias a este tipo de ocorrências. Por isso, o seguro de caçador é um seguro de responsabilidade civil obrigatório e a sua importância acaba por ser relevante dada a quantidade de caçadores existentes em Portugal (Gilberto, 2012). Para o exercício da caça, os caçadores devem celebrar um contrato de seguro de responsabilidade civil destinado a cobrir eventuais danos causados a terceiros, no montante mínimo de 100 000€ (Artigo 76º, Decreto-Lei n.º 24/2018, s.d.). Também os proprietários das Embarcações de Recreio são obrigados a celebrar um contrato de seguro que garanta a responsabilidade civil por danos causados a terceiros (Decreto-Lei n.º 84/2019, s.d.) que se aplica em todo o território nacional, abrangendo a zona económica exclusiva, o mar territorial e as águas interiores portuguesas. O capital mínimo, obrigatório para este seguro, por acidente ou séries de acidentes resultantes do mesmo evento é de 249 398,94€ ou 50 milhões de escudos, seja qual for o número de vítimas ou a natureza dos danos (Portaria n.º 689/2001, s.d.). De seguida, temos o seguro de empregada doméstica onde a entidade empregadora deve transferir a responsabilidade pela reparação dos danos emergentes de acidente de trabalho para entidades legalmente autorizadas a fazer este seguro (Artigo 26º, Decreto-Lei n.º 235/92, s.d.). Para terminar, temos o seguro de incêndio. Apesar de os seguros multirriscos oferecerem soluções que vão para além do necessário, apenas é obrigatório o seguro contra o risco de incêndio do edifício, quer quanto às frações autónomas, quer relativamente às partes comuns (Artigo 1429.º, Decreto-Lei n.º 267/94, s.d.). Essencialmente, as partes comuns são as seguintes partes do edifício: o telhado ou os terraços de cobertura, ainda que destinados ao uso de qualquer fração; as instalações gerais de água, eletricidade, aquecimento, ar condicionado, gás, comunicações e semelhante; as garagens e outros lugares de estacionamento, entre outras (Artigo 1421º, Decreto-Lei n.º 267/94, s.d.).

4.1.1. Tabela de mapeamentos de seguros e/ou coberturas não obrigatórios

Após recolher a informação sobre a obrigatoriedade dos seguros, é necessário perceber como é que se pode transpor este tipo de informação para o modelo, de forma a distinguir quais os clientes que apenas se importam em ter as coberturas obrigatórias dos que são mais sensíveis ao tema e preferem adquirir para além do obrigatório. Para isso, foram analisadas todas as fichas de produtos em vigor da Fidelidade e dos produtos descontinuados, mas que ainda têm apólices ativas. Ao explorar as mesmas, percebeu-se que não era possível isolar as coberturas obrigatórias das outras, visto que as opções mais básicas de cada produto contêm mais do que a cobertura obrigatória. Por exemplo, ao consultar a oferta mais recente da LOB Automóvel no *site* da Fidelidade (Auto 1234, s.d.) é possível ver a diferença entre os quatro produtos:





	 AUTO 1	 AUTO 2	 AUTO 3	 AUTO 4
COBERTURAS				
Viagem legal e bem protegido				
Responsabilidade Civil	Mínimo Obrigatório	Incluído - Mínimo Obrigatório Opcional - 50.000.000€	Incluído - Mínimo Obrigatório Opcional - 50.000.000€	Incluído - Mínimo Obrigatório Opcional - 50.000.000€
Responsabilidade Civil Carga	-	Opcional	Opcional	Opcional
Proteção Vital do Condutor	-	Opcional - 500.000€	Opcional - 500.000€	Opcional - 500.000€
Proteção ao Condutor	10.000€ morte ou invalidez permanente / 1.000€ despesas de tratamento	Opcional - 10.000€ morte ou invalidez permanente / 1.000€ despesas de tratamento	Opcional - 10.000€ morte ou invalidez permanente / 1.000€ despesas de tratamento	Opcional - 10.000€ morte ou invalidez permanente / 1.000€ despesas de tratamento
Proteção Jurídica	-	Opcional - Nível 3	Nível 3	Nível 3
Proteja o seu veículo				
Furto ou Roubo, Fenómenos da Natureza, Incêndio Raio ou Explosão, Atos de Vandalismo	-	-	Capital do Veículo	Capital do Veículo
Choque, Colisão ou Capotamento	-	-	-	Capital do Veículo
Benefício de serviços que o apoiam, em caso de imprevisto				
Assistência em Viagem	Basic	Plus	Plus	Plus
Ocupantes da Viatura	-	10.000€ morte ou invalidez permanente / 1.000€ despesas de tratamento	10.000€ morte ou invalidez permanente / 1.000€ despesas de tratamento	Opcional - 10.000€ morte ou invalidez permanente / 1.000€ despesas de tratamento
Quebra Isolada de Vidros	-	1.000€	1.000€	1.000
Viatura de Substituição	-	-	-	Nível 1

Figura 10 – Parte das coberturas da oferta automóvel no *site* da Fidelidade

Na figura 7 é possível visualizar a oferta mais básica destes produtos AUTO, que para além da Responsabilidade Civil que é obrigatória, tem também a cobertura de Proteção ao condutor e Assistência em Viagem. Este exemplo serve para mostrar que é necessário criar uma tabela para cada LOB obrigatória que mapeie todas as coberturas dos produtos com apólices em vigor, para que desta forma se consiga distinguir as ofertas base, que serão consideradas como obrigatórias, das restantes. De seguida, a Tabela 1 ilustra um exemplo destas mesmas tabelas, onde está demonstrado a oferta Fidelidade *Pets* com a cobertura de responsabilidade civil. Através da variável “Flag Não Obrigatório” vê-se que esta cobertura apenas é obrigatória (“Flag Não Obrigatório”=0) para animais potencialmente perigosos com um capital de 50 000€.

Produto	Cobertura	Índice Perigosidade	Intervalo Capital	Flag Não Obrigatório
Fidelidade PETS	Responsabilidade Civil	1-Animal Companhia	50 000	1
Fidelidade PETS	Responsabilidade Civil	1-Animal Companhia	>50 000	1
Fidelidade PETS	Responsabilidade Civil	2-Potencialmente Perigoso	50 000	0
Fidelidade PETS	Responsabilidade Civil	2-Potencialmente Perigoso	>50 000	1

Tabela 1 - Exemplo da tabela de mapeamento da oferta Fidelidade *Pets*

4.2. MODELO DE PROPENSÃO À COMPRA DE PRODUTOS E/OU SEGUROS NÃO OBRIGATÓRIOS

4.2.1. Identificação do universo e da *target*

Com o mapeamento dos produtos e/ou coberturas obrigatórias e antes do desenvolvimento do modelo, é necessário criar o universo de clientes pretendidos e a respetiva variável *target*. Assim, o universo inclui todos os clientes ativos no mês de Novembro de 2018 que compraram algum produto das seis LOBS referidas anteriormente, no mês seguinte, Dezembro de 2018, tendo sido aplicados os seguintes filtros:

Incluir

- ✓ Pessoas Singulares;
- ✓ Clientes que sejam Tomadores ou Tomadores + Pessoa Segura;
- ✓ Clientes que tenham comprado pelo menos um produto nas seguintes LOBS: Automóvel, Incêndio, Embarcações, Responsabilidade Civil, Caça e Acidentes de Trabalho.

Excluir

- ✗ Clientes falecidos;
- ✗ Clientes bloqueados;
- ✗ Clientes sem incidências;
- ✗ Clientes confidenciais;
- ✗ Clientes a não contactar (destacados pelo *Marketing*).

Para construir a variável *target* olhou-se para os clientes elegíveis, ou seja, os clientes que respeitavam os filtros acima mencionados, juntando-se posteriormente a informação do número de unidades/objetos de risco e/ou pessoas seguras, e o número de coberturas obrigatórias e não obrigatórias por apólice. Após a junção da informação, foram calculados três rácios:

1. Rácio de não obrigatoriedade de cada apólice na LOB

$$R_i^k = \frac{\sum_{j=1}^J \mathbb{1}_{NO} F_{NO_{ij}}}{\sum_{j=1}^J \mathbb{1}_{NO} F_{NO_{ij}} + \sum_{j=1}^J \mathbb{1}_O F_{O_{ij}}}$$

Equação 1 - Rácio de não obrigatoriedade de cada apólice na LOB

Onde R_i^k é o rácio de não obrigatoriedade de cada apólice i na LOB k que é calculado através da divisão entre o somatório das coberturas $j \in \{i, \dots, J\}$ não obrigatórias NO , definidas através de uma *flag* $F_{NO_{ij}}$, e a soma entre o somatório anterior e o somatório das coberturas $j \in \{i, \dots, J\}$ obrigatórias O , definidas através de uma *flag* $F_{O_{ij}}$.

2. Rácio de não obrigatoriedade do cliente na LOB

$$L_u^k = \frac{\sum_{j=1}^J \mathbb{1}_{NO} F_{NO_{uj}}}{\sum_{j=1}^J \mathbb{1}_{NO} F_{NO_{uj}} + \sum_{j=1}^J \mathbb{1}_O F_{O_{uj}}}$$

Equação 2 - Rácio de não obrigatoriedade do cliente na LOB

Seja L_u^k o rácio de não obrigatoriedade do cliente u na LOB k que é calculado pela divisão entre o somatório das coberturas $j \in \{i, \dots, J\}$ não obrigatórias NO , definidas através de uma *flag* $F_{NO_{uj}}$, e a soma entre o somatório anterior e o somatório das coberturas $j \in \{i, \dots, J\}$ obrigatórias O , definidas pela *flag* $F_{O_{uj}}$.

3. Rácio de não obrigatoriedade total do cliente

$$C_u = \frac{\sum_{j=1}^J \mathbb{1}_{NO} F_{NO_{uj}}}{\sum_{j=1}^J \mathbb{1}_{NO} F_{NO_{uj}} + \sum_{j=1}^J \mathbb{1}_O F_{O_{uj}}}$$

Equação 3 - Rácio de não obrigatoriedade total do cliente

Onde C_u é o rácio de não obrigatoriedade total do cliente u que é calculado pela divisão entre o somatório das coberturas $j \in \{i, \dots, J\}$ não obrigatórias NO , definidas através de uma *flag* $F_{NO_{uj}}$, e a soma entre o somatório anterior e o somatório das coberturas $j \in \{i, \dots, J\}$ obrigatórias O , definidas pela *flag* $F_{O_{uj}}$.

O último rácio foi a base para a definição da variável *target* e, como tal, é necessário olhar para a distribuição do mesmo.

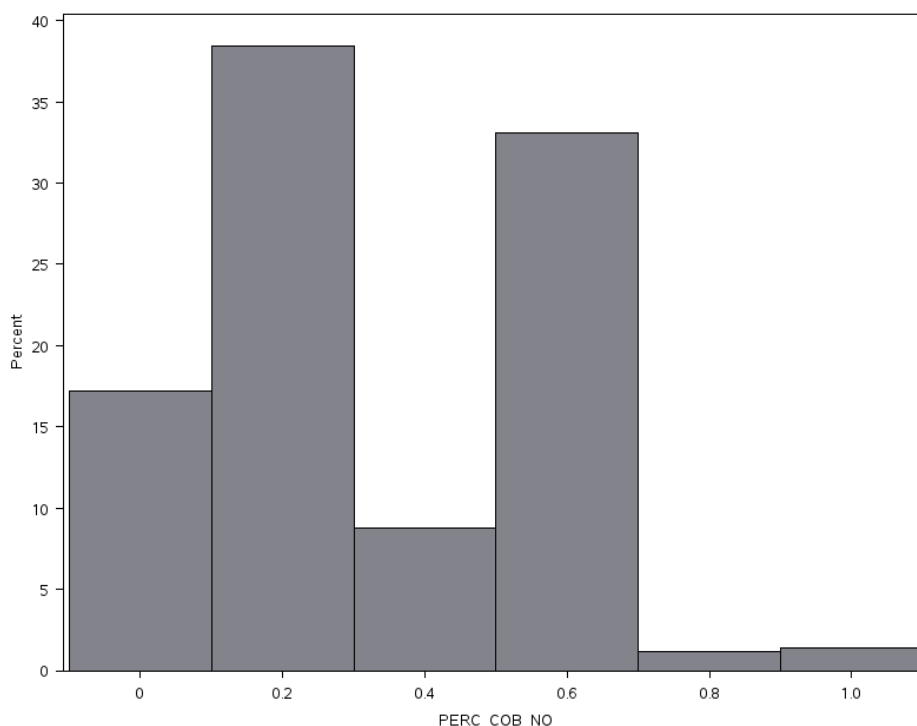


Figura 11 – Distribuição do rácio de não obrigatoriedade total

Na figura 11 acima apresentada, é possível verificar que os picos da distribuição estão entre os intervalos $[0,1; 0,3]$ e $[0,5; 0,7]$. Como não faz sentido definir uma *target* onde a propensão de não obrigatoriedade é superior a 50%, o que significaria que os clientes teriam metade das suas coberturas como obrigatórias, para além de que a maior concentração de clientes está situada na zona oposta do gráfico, o foco irá para o intervalo de $[0,1; 0,3]$.

Assim, os clientes com um rácio de não obrigatoriedade total superior a 0,20 foram classificados como *Target=1* e os clientes com um rácio inferior ou igual estão identificados como *Target=0*. Com a *target* definida, observa-se o mês de análise e verifica-se que a taxa de propensão a adquirir coberturas e/ou produtos não obrigatórios é muito baixa, tal como a tabela 2 demonstra e, por isso, para finalizar a construção do universo do modelo é necessário ampliar a janela temporal do mesmo.

Dezembro/2018	
Nº Clientes Elegíveis	1 872 695
Nº Clientes taxa > 20% de coberturas não obrigatórias	6 224
Taxa de propensão	0,0332 %

Tabela 2 – Taxa de propensão do universo no mês de análise

O universo final do modelo é constituído por todos os clientes ativos da Fidelidade desde **Janeiro de 2018** até **Dezembro de 2018** que não tinham nenhum produto no mês anterior, das seis LOBS referidas, e que compraram no mês seguinte. Na tabela 3 é possível ver a nova taxa de propensão do universo.

Jan/2018 - Dez/2018	
Nº Clientes Elegíveis	1 872 695
Nº Clientes taxa > 20% de coberturas não obrigatórias	62 627
Taxa de propensão	3,34%

Tabela 3 – Taxa de propensão do universo final

4.2.2. Preparação e tratamento de dados

Depois da identificação e construção do universo e antes da modelação é necessário explorar os dados que irão ser utilizados. A tabela do universo inicial é constituída por 1385 variáveis, porém é possível excluir variáveis *a priori*. Assim, foram retiradas 266 variáveis que correspondem a informação do cliente, como o número de telefone, a morada, as coordenadas geográficas da mesma, entre outras. Foram também retiradas variáveis relativas ao exercício atual, visto que são variáveis voláteis e não comparáveis entre clientes, por exemplo, em termos de prémios apenas podemos comparar clientes pelo seu PCA, pois sabemos que aquele valor é estático e que corresponde ao valor total pago pelo tomador à seguradora, isto se não acrescentar mais nenhuma cobertura ou não contratar mais nenhum seguro. Já os prémios adquiridos no exercício atual são valores que já foram pagos até ao momento em que se está a analisar, e estes valores não se podem comparar entre clientes, pois dependem do início de cada apólice, dos sinistros que já tiveram, dos custos dos mesmos, etc.

De seguida, analisou-se a presença de valores omissos onde as variáveis que contêm 30% ou mais são eliminadas do universo, resultando numa redução de 183 variáveis. Na tabela 4 estão representadas o número de variáveis por intervalo de percentagem de valores omissos.

Intervalo de percentagem	Nº de variáveis
[100% – 80%]	160
]80% – 50%]	11
]50% – 30%]	12

Tabela 4 – Nº variáveis por intervalo de percentagem de valores omissos

Passando para as variáveis categóricas, é importante avaliar aquelas que apresentam demasiados níveis distintos, pois podem causar enviesamento no resultado final do modelo. Para maximizar a capacidade de discriminação do modelo as variáveis com mais de 10 níveis foram sujeitas a uma transformação. Assim, para estas variáveis, ficaram apenas as 10 classes com mais expressão juntando as restantes numa categoria que se denomina de “Outros”, a figura 12 exemplifica uma dessas modificações. As variáveis categóricas que sofreram a alteração descrita foram as seguintes: descrição do cabaz atual, descrição do cabaz há 12 meses, descrição da família de produtos da última apólice comprada, descrição da família da apólice em vigor mais antiga, descrição da família da apólice em vigor mais recente, descritivo da família de produtos do mediador preferencial com maior PCA,

descrição da família com mais apólices na carteira do mediador preferencial, descrição do meio de pagamento preferencial e descrição do mediador/agente preferencial.

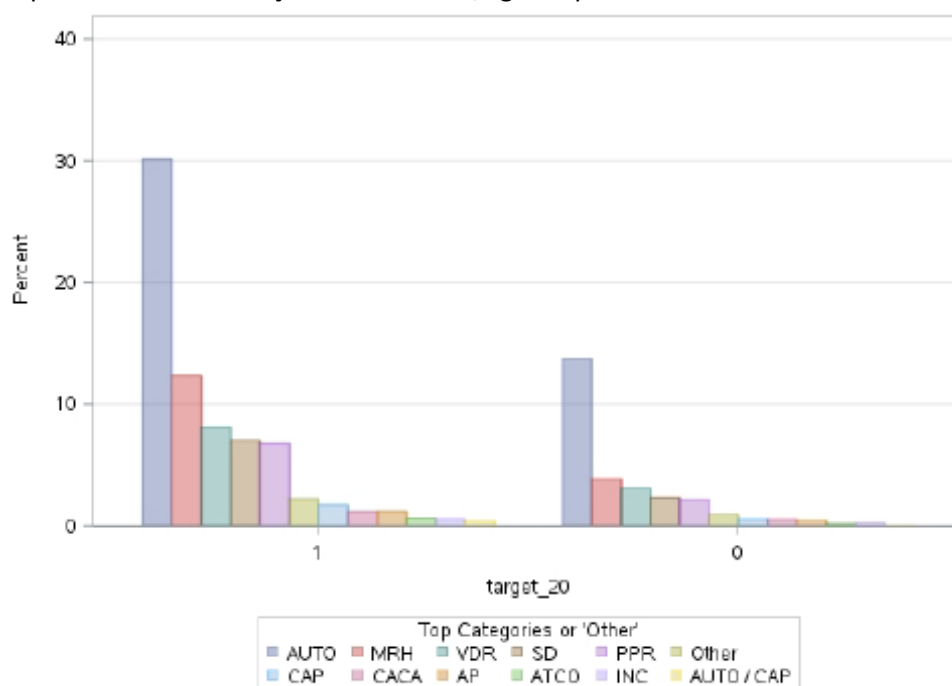


Figura 12 – Recategorização da variável Descrição da família da apólice em vigor mais antiga

De forma a encontrar padrões nas variáveis em relação à *target*, foi gerada uma série de gráficos *bar chart* que demonstram o seu comportamento nas classes das variáveis categóricas ou nos valores das variáveis numéricas. Assim, foram eliminadas as variáveis que não permitem discriminar os clientes propensos à compra de seguros e/ou coberturas não obrigatórias (*target*=1) dos restantes (*target*=0). A figura 13 é um exemplo de uma variável não discriminante, a variável *Flag Adulto*, que dada a aplicação do filtro “Clientes que sejam Tomadores ou Tomadores + Pessoa Segura” faz com que apenas sejam selecionados para o universo, clientes com idade igual ou superior a 18 anos.

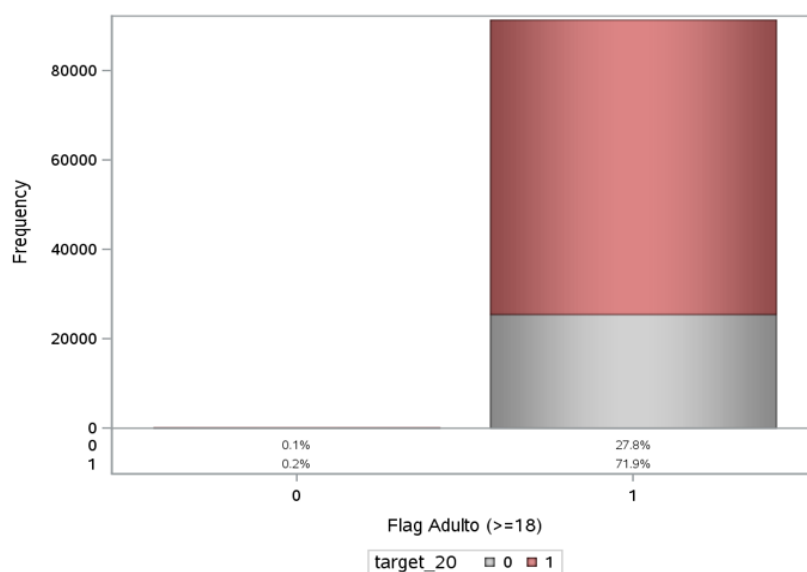


Figura 13 - Gráfico de frequência da variável *Flag Adulto*

No entanto, variáveis com elevada capacidade discriminatória em relação à *target* foram mantidas, como é o caso da variável Quantidade de meses para expirar/renovar a próxima apólice Automóvel, na figura 14, onde se verifica que a probabilidade de um cliente comprar um seguro e/ou produto não obrigatório (*target*=1) aumenta consideravelmente quando faltam 12 meses para expirar ou renovar a sua próxima apólice automóvel. Apesar desta LOB ser obrigatória, a maioria dos clientes que possui uma apólice automóvel não adquire apenas as coberturas obrigatórias por lei, a cobertura de quebra isolada de vidros é um excelente exemplo disso.

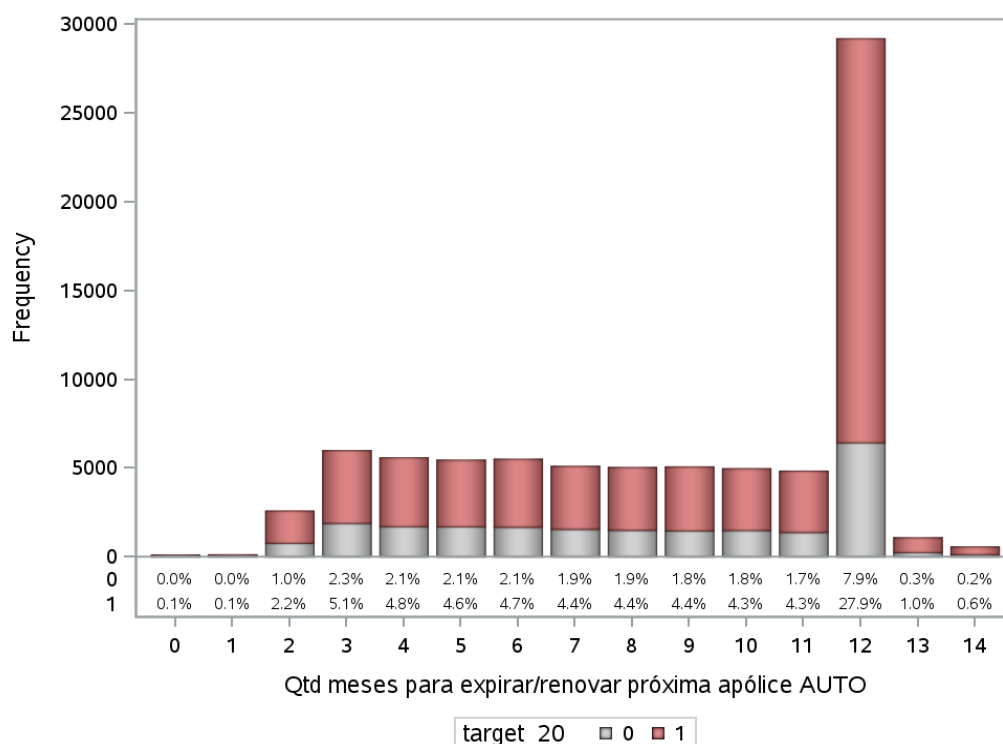


Figura 14 - Gráfico de frequência da variável “Quantidade de meses para expirar a próxima apólice Automóvel”

4.2.3. Pré-Processamento dos dados

Após uma análise exaustiva dos dados, e consequente tratamento dos mesmos, avança-se para uma fase que tem como objetivo acrescentar qualidade ao modelo. Começando por realizar o particionamento dos dados em dois *datasets*: o primeiro contém 70% dos dados e corresponde ao conjunto de Treino, e o segundo é o conjunto de Validação com os restantes 30%. Esta divisão garante que o treino seja feito de forma representativa para que não haja problemas de *overfitting*. O método de particionamento ficou como *Default*, o que significa que o *Simple Random* foi selecionado. Neste método cada observação no *dataset* tem a mesma probabilidade de ser gravada num dos conjuntos particionados (SAS, Data Partition Node, s.d.). A tabela 5 ilustra o número de observações que cada *dataset* contém.

Dataset	Nº de observações
Train	60 622
Validate	25 983

Tabela 5 – Particionamento da tabela

De forma a reduzir a dimensionalidade dos dados através da observação da correlação entre variáveis, remove-se variáveis redundantes e irrelevantes para o modelo. Assim, utilizando o nó *Variable Clustering* tem-se acesso à matriz de correlação de *Pearson*, onde este coeficiente mede a força das relações lineares entre duas variáveis e o seu valor está entre o intervalo de -1 a 1. Quanto mais próxima estiver a correlação dos extremos do intervalo, mais forte será a associação entre variáveis, seja ela uma associação positiva, próxima de 1, ou negativa, próxima de -1. O coeficiente de *Pearson* tenta desenhar uma linha de *best fit* entre as duas variáveis em estudo, o que significa que quanto mais próxima estiver a correlação dos extremos menos variação existe em torno da linha desenhada, e quanto mais próxima estiver a correlação do valor 0 maior a sua variação (statistics, s.d.).

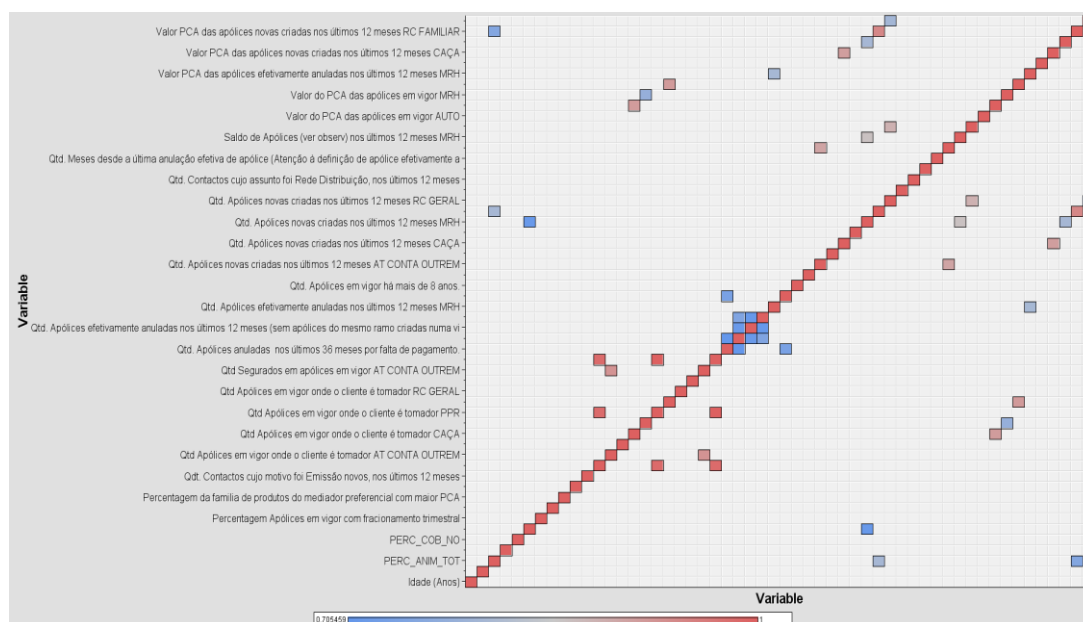


Figura 15 – Matriz de correlação de *Pearson* filtrada

Com o intuito de excluir as associações mais fortes, foram eliminadas as variáveis com uma correlação igual ou superior ao valor absoluto de 0,70. Na figura 15 é visível a aplicação do filtro, indicando assim quais as variáveis correlacionadas entre si.

Com esta última análise reduz-se o *dataset* que irá alimentar o modelo para 62 variáveis, distribuídas pelas categorias listadas na tabela 6:

Categoria	Nº de Variáveis
Target e identificação do cliente	3
Informação do cliente: sexo, idade, etc.	5
Informação geográfica: NUTS, etc.	3
Informação do mediador e método de pagamento	11
Informação do cabaz: quantidade de apólices, maturidade, saldo, etc.	40

Tabela 6 – Categorias das variáveis *input*

4.2.2.1. Métodos de seleção de variáveis

Gradient Boosting

Este método utiliza um algoritmo de partição que procura a partição ótima dos dados para uma única variável *target*. É uma abordagem que faz a reamostragem dos dados em análise para gerar resultados que formam uma média ponderada do *dataset* reamostrado. Uma árvore *boosting* cria uma série de árvores de decisão que formam um único modelo preditivo. Tal como as árvores de decisão, o *gradient boosting* não tira pressupostos sobre a distribuição dos dados. O *boosting* é menos propenso ao sobreajustamento dos dados do que uma única árvore de decisão. Se uma árvore de decisão se ajusta bastante bem aos dados, então o *gradient boosting* irá melhorar esse ajuste (SAS, Create a Gradient Boosting Model of the Data, s.d.).

Variable Clustering

É uma ferramenta bastante útil para a redução dos dados porque encontra as melhores variáveis para a análise, retirando a colinearidade, diminuindo a redundância de variáveis e ajudando a revelar a estrutura subjacente das variáveis de *input* no *dataset* (SAS, Overview of Variable Clustering, s.d.). Este método divide as variáveis numéricas em *clusters* hierárquicos ou disjuntos, baseando-se numa métrica de similaridade. Os *clusters* são criados para que as variáveis do mesmo *cluster* sejam altamente correlacionadas entre si mas que tenham uma baixa correlação com outro qualquer *cluster*. O processo de seleção consiste em selecionar uma variável de cada *cluster*, variável essa que tem de ter o mínimo *R-Square Ratio* dentro do próprio *cluster*. A variável selecionada é a que melhor representa melhor o seu *cluster* (Aggarwal & Kossian, 2011).

$$R\text{ Square Ratio} = \frac{1 - R\text{ Square Own Cluster}}{1 - R\text{ Square Next Closest}}$$

Equação 4 – Cálculo do *R Square Ratio*

Variable Selection

Este nó fornece um método alternativo para eliminar variáveis irrelevantes e selecionar as variáveis que tenham maior poder preditivo. No processo de seleção existem dois métodos disponíveis, o R-Quadrado e o Qui-Quadrado. O método do R-Quadrado pode ser utilizado para *targets* binárias e intervalares, e é realizado em duas fases: na primeira, é calculado o R-Quadrado entre as variáveis *input* e *target*, onde todas as variáveis com uma correlação superior ao limite definido (como *default* está 0,005) são selecionadas para a segunda fase. Na segunda fase é utilizado o processo sequencial de seleção *Forward*, que começa por selecionar a variável *input* com o maior coeficiente de correlação com a *target*. Um modelo de regressão é estimado com o *input* selecionado, e a cada etapa, a variável *input* que fornece a maior contribuição incremental para o modelo R-Quadrado é adicionada à regressão. Quando o limite inferior dessa contribuição for atingido, o processo de seleção para (Sarma & Corp, 2007). O método Qui-Quadrado apenas pode ser utilizado para *targets* binárias e fornece uma rápida avaliação preliminar das variáveis, facilitando o desenvolvimento rápido de modelos preditivos com grandes volumes de dados. A seleção Qui-Quadrado usa a divisão da variável binária para maximizar os valores do Qui-Quadrado de uma tabela de frequência 2 x 2 (SAS, Variable Selection Node, s.d.).

Regressão

A regressão pode selecionar o modelo de um conjunto de termos, denominados também de efeitos. Por exemplo, um modelo contém, para além do *intercept*, os efeitos A, B, A*B e A*A. Os efeitos A e B são conhecidos como os principais, o A*B é chamado de efeito de interação e o A*A é o efeito polinomial. No nó da regressão, pode-se especificar o conjunto de efeitos candidatos, tais como os quatro acima mencionados, e o nó pode selecionar o modelo a partir desses candidatos. Em cada fase do método de seleção, um efeito é adicionado ou eliminado do modelo para que uma sequência de modelos seja gerada, uma em cada etapa. Pode-se selecionar um dos três modelos de seleção de efeitos (SAS, Regression Node, s.d.): *Backward*, *Forward* e *Stepwise*.

O *Backward*, por defeito, começa com todos os efeitos candidatos do modelo e vai eliminando sistematicamente os que não são significativamente associados com a *target*, até atingir o nível de significância de permanência, definida por *default* como 0,05, ou o número de variáveis de paragem. Seguindo para o método *Forward*, que começa sem nenhum efeito candidato no modelo e adiciona sistematicamente os efeitos que são significativamente associados com a *target* até que nenhum dos efeitos restantes respeite o nível de significância de entrada, definido por *default* como 0,05, ou seja atingido o número de variáveis de paragem. Por fim, o método *Stepwise*, que é semelhante ao método *Forward*, começa sem nenhum dos efeitos candidatos no modelo e adiciona sistematicamente os efeitos que são significativamente associados com a *target*. Porém, depois de um efeito ser adicionado ao modelo, este método pode retirá-lo mais tarde se o mesmo deixar de estar significativamente associado com a *target*. Este processo acaba quando atingir o nível de significância de permanência, o número máximo de etapas definido pelo utilizador ou quando um efeito adicionado numa etapa é o eliminado na etapa seguinte (SAS, Regression Node, s.d.).

4.2.4. MODELAÇÃO

Após toda a análise, tratamento e seleção de variáveis, chega-se assim à fase de implementação de algoritmos de *machine learning* que irão prever a propensão à compra de produtos e/ou seguros não obrigatórios. Pegando no ponto anterior da seleção de variáveis, são aplicados para cada método quatro tipos de modelos preditivos: *Random Forest*, Regressão Logística, Árvores de Decisão e Rede Neuronal. Na figura 16 é possível ver o diagrama simplificado do modelo desenvolvido em SAS Miner, onde estão representados os melhores modelos por cada método de seleção de variáveis, isto para que o diagrama não se torne demasiado extenso e confuso.

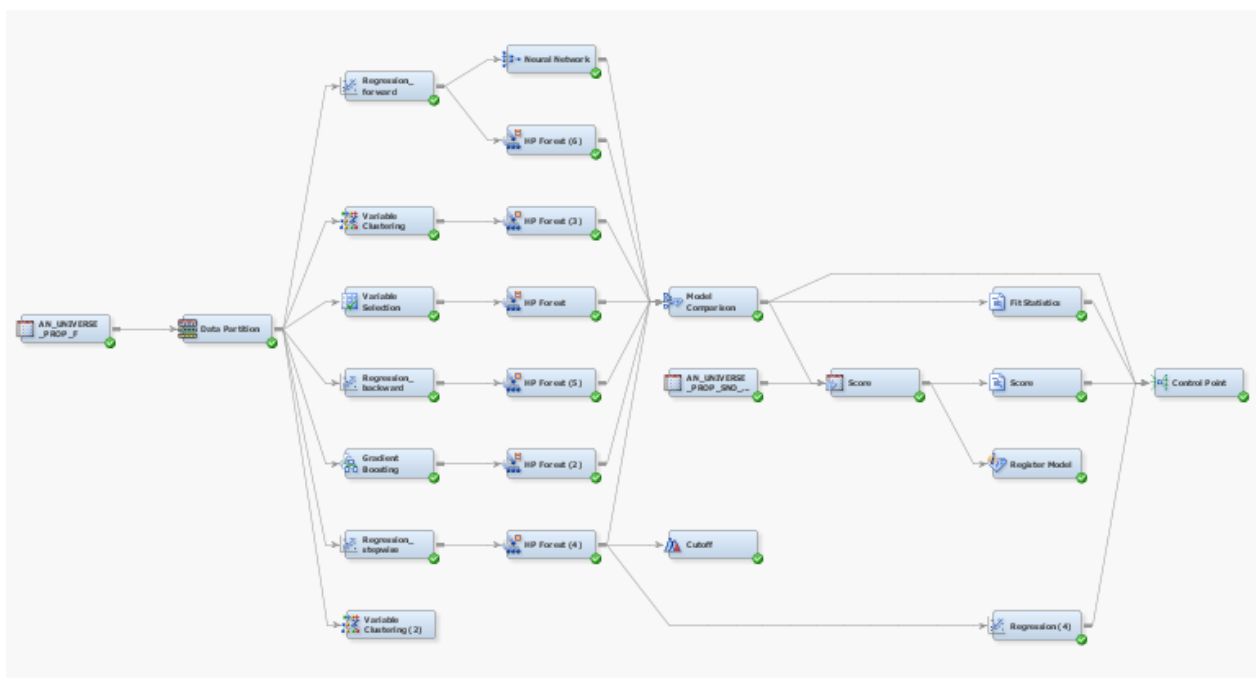


Figura 16 – Diagrama do modelo em SAS Miner

De forma a perceber qual o algoritmo que apresenta o melhor resultado, é utilizado o nó *Model Comparison* que realiza a comparação dos diversos modelos e, de forma automática, escolhe o modelo *champion* através dos critérios definidos por *default* ou pelo utilizador. Neste caso, o modelo *champion* será aquele que terá um melhor valor para a métrica *Lift*. Apesar de a escolha ser realizada de forma automática é efetuada uma análise a todas as estatísticas utilizadas para comparação. Na tabela 7 está presente a demonstração de resultados, onde apenas serão exibidas algumas estatísticas dos quatro melhores modelos para que a interpretação seja mais fácil.

	HP Forest 6	HP Forest 5	HP Forest 4	HP Forest 2
ROC Index				
Train	0,886	0,881	0,886	0,879
Validate	0,872	0,866	0,873	0,866
Lift				
Train	1,3669	1,3651	1,3674	1,3705
Validate	1,3605	1,3541	1,3605	1,3520
Lift Cumulative				
Train	1,3731	1,3719	1,3737	1,3747
Validate	1,3685	1,3637	1,3696	1,3653
Cumulative % Response				
Train	99,291	99,208	99,340	99,406
Validate	98,961	98,615	99,038	98,730

Tabela 7 – Estatísticas dos quatro melhores modelos – Treino e Validação

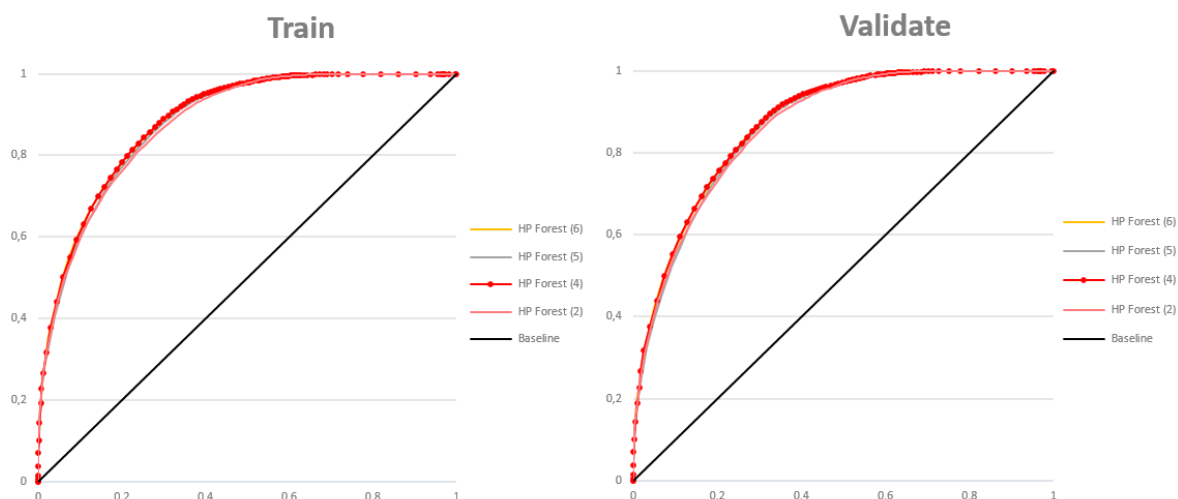


Figura 17 - Curva ROC: Treino e Validação

Através da tabela 7 e dos gráficos apresentados na figura 17, é possível ver que os quatro melhores modelos são todos *Random Forest*, porém o modelo *champion* é o “*HP Forest 4*” que tem como base o método de Regressão *Stepwise* para a seleção de variáveis. Apesar deste modelo ser o segundo melhor nos dados de treino, é o que apresenta melhores resultados nos dados de validação e, por isso, é o modelo escolhido. O facto de se dar mais importância às métricas obtidas nos dados de validação é para que se consiga evitar problemas de *overfitting*, pois obtém-se uma estimativa demasiado otimista ao calcular o erro generalizado do modelo nos mesmos dados utilizados para escolher o modelo *champion* (Raschka, 2018).

4.2.5. RESULTADOS

Concluindo o desenvolvimento do modelo de propensão, é importante analisar os resultados do mesmo. A tabela 8 representa a distribuição das 22 variáveis finais, sem contar com a *target*, por diversas categorias de informação. Onde se destaca a informação do cabaz do cliente com 14 variáveis. Nesta categoria estão presentes variáveis que dão informação sobre o cabaz atual e homólogo do cliente, a quantidade de apólices novas e anuladas no ramo automóvel, entre outras.

Categoria	Nº de Variáveis
<i>Target</i> e identificação do tipo de cliente	3
Informação do cliente: sexo, idade, etc.	2
Informação geográfica: NUTS, etc.	1
Informação do mediador e método de pagamento	3
Informação do cabaz: quantidade de apólices, maturidade, saldo, etc.	14

Tabela 8 – Categorias das variáveis selecionadas pelo modelo final

Por forma a obter uma análise mais profunda sobre os resultados, é necessário interpretar o modelo. Quanto maior for a interpretabilidade de um modelo, mais fácil será para os utilizadores compreenderem porque certas decisões ou previsões foram feitas (Molnar). Os modelos de *machine learning* podem ser classificados em duas categorias: os modelos *white box*, que incluem as árvores de decisão e modelos de regressão, e os modelos *black box* como *random forests*, *gradiente boosting* e redes neuronais (Olaleye, 2019). Assim, para fornecer um entendimento global sobre o que foi aprendido pelo modelo, mede-se a importância das variáveis, que indica a contribuição estatística de cada variável ao modelo subjacente (Güneş, Tharrington, Abbey, & Hunt, 2020). A figura 18 apresenta os valores da métrica *Gini* que mede a importância de cada variável, importância essa que é proporcional à soma da redução da impureza do nó, somada aos nós que a variável divide (SAS, The HPFOREST Procedure, s.d.). Assim, as variáveis que mais contribuem para o modelo destacam-se pelo facto de estarem associadas a novas apólices compradas e nomeadamente apólices compradas no ramo Automóvel. Tal como dito anteriormente, o ramo Automóvel é a LOB com mais expressão na carteira de qualquer seguradora portuguesa, o que pode explicar a sua elevada contribuição para este modelo.

Variável	GiniOOB
PCA das apólices novas criadas nos últimos 12 meses AUTO	0,06398
Família de produtos da última apólice comprada	0,01926
Família da apólice em vigor mais recente	0,01679
Quantidade de apólices novas criadas nos últimos 12 meses no ramo Automóvel	0,01116
Descrição do Mediador / Agente Preferencial	0,00664
Descrição NUT2	0,00451
Saldo de Apólices nos últimos 12 meses no ramo Multirriscos Habitação	0,0036
Quantidade de apólices em vigor onde o cliente é tomador Automóvel;	0,00345
Cabaz atual	0,00334
Quantidade de apólices em vigor há menos de 2 anos;	0,00218
Segmento marketing do cliente	0,00185
Sexo	0,00095
Idade	0,00083
Família da última anulação efetiva da apólice	0,00072
Estado do cliente	0,00062
Quantidade de apólices efetivamente anuladas nos últimos 12 meses no ramo Automóvel	0,00057
Quantidade de apólices em vigor onde o cliente é tomador Responsabilidade Civil Familiar	0,00036
Quantidade de apólices novas criadas nos últimos 12 meses no ramo Incêndio	0,0003
Fracionamento de pagamento preferencial	0,00028
Cabaz há 12 meses	0,00024
Quantidade de apólices em vigor há mais de 8 anos	0,00014
Percentagem de apólices em vigor com fracionamento único	-0,00008

Figura 18 – Importância das variáveis no modelo final

Para além da contribuição preditiva das variáveis, é também possível analisar se todas as variáveis são relevantes e perceber quais os impactos positivos ou negativos para o resultado final do modelo. Um dos métodos possíveis de interpretar um modelo *black box* como o modelo *random forest* selecionado, é através da utilização de um modelo inerente interpretável, também chamado de modelo explicável, que incorpora a interpretabilidade diretamente na estrutura do modelo sendo, portanto, auto-explicativo (Güneş, Tharrington, Abbey, & Hunt, 2020). Recorrendo a uma Regressão Logística, analisa-se os parâmetros das variáveis e verifica-se que cada variável é

estatisticamente significativa para o modelo, uma vez que, a probabilidade obtida no teste do Qui Quadrado é inferior ao valor do *p-value* (0,05). De seguida, analisa-se os valores da *odds ratio* que consiste numa medida de associação utilizada para comparar as probabilidades relativas de ocorrer um determinado resultado mediante as variações da variável explicativa (ou de interesse). Através desta medida é possível determinar o valor do impacto que cada variável explicativa exerce sobre a variável dependente, a variável *target*, através do seguinte cálculo (SAS, The LOGISTIC Procedure, s.d.):

$$(odds\ ratio - 1) \times 100$$

Equação 5 – Cálculo da medida *odds ratio*

Na figura 19 estão representados alguns dos efeitos das variáveis selecionadas, sendo que os restantes podem ser encontrados na seção “Anexos”. Posto isto, a medida *odds ratio* pode ser interpretada da seguinte forma: quando um cliente tem um aumento de 1 unidade na variável “Quantidade de apólices em vigor há menos de 2 anos”, a probabilidade de propensão à compra de coberturas e/ou seguros obrigatórios diminui 4,9%, por exemplo. No caso da interpretação das variáveis categóricas, a leitura realiza-se da seguinte maneira: a probabilidade de propensão de um cliente que pertença à zona Centro, variável “Descrição da NUT 2”, é 27,2% superior quando comparada com a probabilidade de propensão de um cliente que viva na Região do Norte.

		Pr > ChiSq	Odds	Effects	
Qtd. Apólices em vigor há menos de 2 anos					
		0.0012	0,951	↓	-4,9%
Descrição da NUT 2					
CENTRO	REGIÃO DO NORTE	<.0001	1,272	↑	27,2%
REGIÃO AUTÓNOMA DA MADEIRA	REGIÃO DO NORTE		0,492	↓	-50,8%
REGIÃO AUTÓNOMA DOS AÇORES	REGIÃO DO NORTE		0,408	↓	-59,2%
REGIÃO DE LISBOA	REGIÃO DO NORTE		0,666	↓	-33,4%
REGIÃO DO ALENTEJO	REGIÃO DO NORTE		1,218	↑	21,8%
REGIÃO DO ALGARVE	REGIÃO DO NORTE		0,893	↓	-10,7%

Figura 19 – Valores do impacto das variáveis “Quantidade de apólices em vigor há menos de 2 anos” e “Descrição da NUT 2”

Concluindo a análise dos resultados, verifica-se que as variáveis dão especialmente ênfase à quantidade de apólices mais recentes, seja em relação à compra ou à anulação. Quanto às LOBs que mais se destacam, são os ramos Automóvel e Multirrisco Habitação, isto porque são ramos dos quais fazem parte os seguros e/ou coberturas obrigatórias mais populares na ótica do consumidor. De notar que em relação ao cliente, as variáveis estão mais viradas para as características do mesmo, como a idade, género e localização geográfica, mas também, caracterizam o seu perfil de consumidor com a descrição do respetivo mediador/agente preferencial e do segmento de marketing no qual se insere.

5. NEXT BEST ACTION – APLICAÇÃO

Na primeira versão do NBA fica assim contemplado o pilar mais completo de toda a solução, o dos Produtos & Serviços, onde estão englobados todos os modelos de *cross-sell* desenvolvidos até então. De seguida, com o coeficiente do ciclo comercial e o coeficiente da especificação de produtos é possível calcular o *score* final que irá permitir calcular e indicar qual a melhor oferta para cada cliente.

Quanto ao coeficiente do ciclo comercial, as LOBs integrantes do mesmo passam a ter um peso de 0,80 (*Commercial Cycle*=“Yes”), ao passo que nas restantes é-lhes atribuído um peso de 0,20 (*Commercial Cycle*=“No”).

O coeficiente da especificação de produtos é atribuído através do modelo de propensão à compra de produtos e/ou seguros não obrigatórios, onde os clientes com um coeficiente superior a 0,20 (*Non Compulsory*=“Yes”) são considerados clientes propensos à compra e, os que têm um coeficiente inferior a 0,20 (*Non Compulsory*=“No”), são clientes que apenas se interessam em ter o que é obrigatório por lei. Para além deste coeficiente, o modelo de propensão também irá atribuir a cada cliente a sua respetiva propensão à compra.

Para exemplificar o cálculo do novo *score* irão ser apresentados dois casos: no primeiro, o cliente terá uma propensão à compra elevada, ou seja, é um cliente propenso a comprar seguros e/ou coberturas não obrigatórias. No segundo caso, o cliente terá uma propensão baixa, o que significa que é um cliente que apenas se interessa em ter a oferta base.

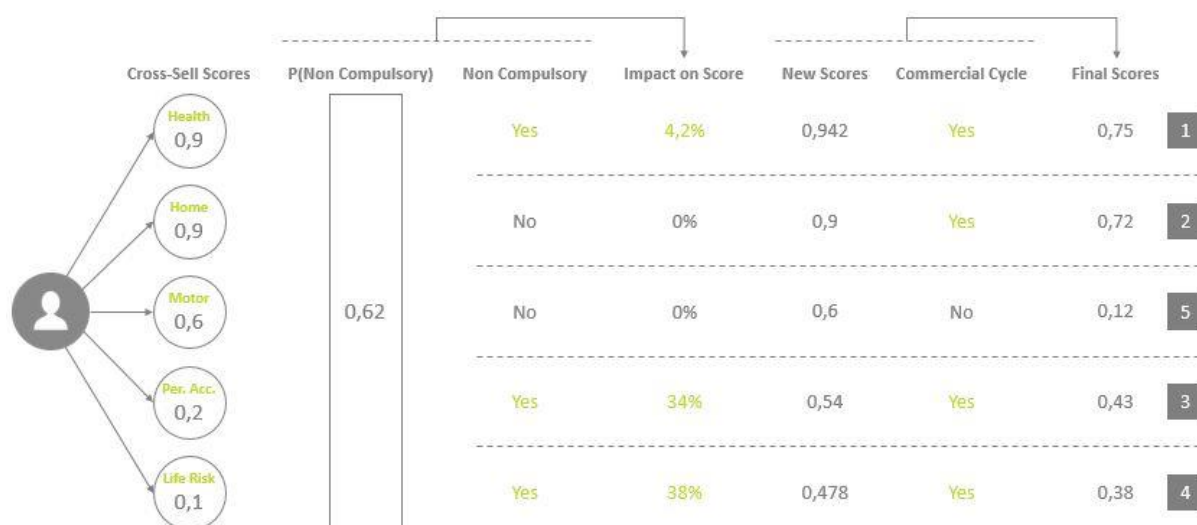


Figura 20 – 1º caso do cálculo do *score* final do NBA

Tal como foi dito anteriormente, no primeiro caso, que está representado pela figura 20, é possível ver um cliente com uma elevada propensão à compra ($P(\text{Non Compulsory}) = 0,62$). Este cliente é elegível para cinco ofertas de *Cross-Sell* com a respetiva indicação se a LOB é obrigatória ou não (*Non Compulsory*). Se a oferta disser respeito a um ramo não obrigatório, e visto que o cliente tem uma elevada propensão para a compra dos mesmos, o impacto no *score* será positivo. No caso dos produtos obrigatórios, o impacto no *score* é nulo.

Com os novos *scores* temporários, é necessário ver se as LOBs estão em ciclo comercial, é neste passo que será aplicado o coeficiente do ciclo comercial que irá tomar o valor de 0,8 caso o

ramo seja não obrigatório e o valor de 0,2 caso contrário. Com todos os ponderadores é finalmente calculado o *score* final para cada ramo, onde é possível ver que a melhor oferta é a de Saúde e, de seguida, de Multiriscos Habitação. Este cálculo permite a ordenação das ofertas de forma a recomendar a melhor oferta para cada cliente, mas também permite desempatar caso haja ramos com o mesmo *score* inicial como era o caso de Saúde e de Multiriscos Habitação.

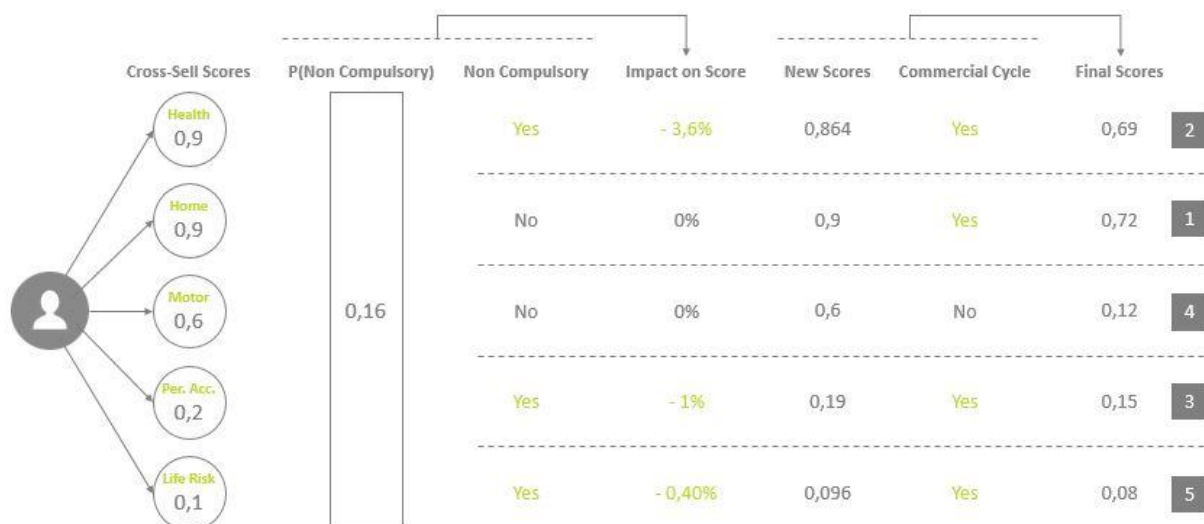


Figura 21 - 2º caso do cálculo do *score* final do NBA

A figura 21 ilustra o segundo caso, onde são apresentadas novamente as cinco ofertas de *Cross-Sell* com a indicação se a LOB é obrigatória ou não. Contudo, desta vez, a propensão do cliente é baixa ($P(\text{Non Compulsory}) = 0,16$). Neste segundo caso, se o produto não for obrigatório, e visto que o cliente tem uma baixa propensão para a compra dos mesmos, o impacto no *score* final será negativo. No caso dos produtos obrigatórios o impacto será novamente nulo. Depois do cálculo dos novos *scores* temporários é aplicado o coeficiente do ciclo comercial e, por fim, são calculados os *scores* finais. Desta vez, a melhor oferta passa a ser a de Multiriscos e, de seguida, a de Saúde.

6. CONCLUSÕES

A inovação no sector dos seguros é um tópico que nos últimos anos se tornou cada vez mais apelativo. Este projeto visa mostrar que é possível aplicar sistemas de recomendação, que recorrentemente são utilizados em setores mais digitais, em setores mais tradicionais como os seguros, mas para tal, é necessário implementar as devidas adaptações ao setor e à informação disponível.

O projeto *Next Best Action* tinha como objetivo implementar um SR que permita obter uma visão global e transversal de cada cliente e, para além disso, fornecer o produto e/ou serviço mais adequado atendendo às suas necessidades.

A solução ideal seria um sistema híbrido *memory-based*, ainda assim, devido à informação disponível de momento na companhia, não é possível aplicar. Por isso será desenvolvido um sistema *model-based*, aproveitando todos os modelos já existentes.

Para a implementação deste projeto foi necessário fazer um levantamento do processo implementado até então. Assim, verificou-se a existência de modelos de *cross-sell*, *up-sell* e *churn*. Com toda a informação recolhida passou-se ao desenho da solução, que consiste num sistema que assenta em três pilares e três coeficientes. O primeiro pilar é o da Retenção, que irá incluir os modelos de *churn*, o segundo pilar é o dos Produtos & Serviços que será alimentado pelos modelos de *cross-sell* e *up-sell* e o terceiro pilar é o das ações, que irá abranger todas as potenciais ações direta ou indiretamente ligadas com os produtos e serviços existentes. Estes pilares serão apoiados pelo coeficiente do ciclo comercial, que identifica quais os produtos que estão em ciclo, pelo coeficiente da retenção, que determina a urgência de um contacto e/ou ação preventiva e, por fim, o coeficiente de especificação de produtos, que transmite o nível de sensibilidade dos clientes ao tema seguros.

Com o plano da solução, passou-se ao desenvolvimento do modelo de propensão à compra de produtos e/ou seguros não obrigatórios que irá alimentar o coeficiente de especificação de produtos. Para a evolução do modelo foi necessário fazer um levantamento dos seguros obrigatórios por lei e construir tabelas que guardassem essa informação para utilização futura. No modelo de propensão foi aplicado a metodologia SEMMA, para organizar todo o processo, e quanto aos modelos de *machine learning*, foram utilizados quatro tipos: Regressão Logística, *Random Forest*, Rede Neuronal e Árvores de Decisão.

O modelo final é uma *Random Forest* que tem como método de seleção de variáveis uma Regressão *Stepwise* e que conta com 22 variáveis finais. Estas variáveis concentram-se essencialmente na informação do cabaz de produtos, nomeadamente na quantidade de apólices mais recentes, sejam elas novas ou anuladas. Para além das quantidades, as LOBs Automóvel e Multiriscos têm também bastante peso na propensão à compra de produtos e/ou seguros não obrigatórios. Para uma análise mais detalhada das variáveis selecionadas, observou-se os valores de Gini, de modo a perceber qual a importância global de cada variável, e para medir o impacto das mesmas, recorreu-se a uma regressão logística.

Por fim procedeu-se à criação da primeira versão do *Next Best Action*, onde ficaram demonstrados dois casos: no primeiro, o cliente tem uma elevada propensão à compra de produtos

e/ou seguros não obrigatórios e, no segundo, o cliente apresenta uma baixa propensão. Para os dois, foi calculado um *score* para todas as ofertas e/ou produtos a que o cliente estivesse elegível. No cálculo deste *score* foram incluídas todas as propensões dos modelos do pilar Produtos & Serviços, bem como os coeficientes do ciclo comercial e da especificação dos produtos.

Em suma, a primeira versão do NBA apesar de não ser a solução planeada inicialmente, tem demonstrado ser uma grande ajuda na altura de contactar o cliente com a proposta de novas ofertas, uma vez que a companhia passou a ter uma visão ordenada sobre as mesmas. Por outras palavras, antes da implementação do NBA a nova oferta era escolhida tendo em conta apenas se estava inserida em ciclo comercial. Se o cliente fosse elegível a mais do que uma oferta em ciclo, escolhia-se a LOB com o maior valor de propensão. Por sua vez, no NBA, tem-se em conta não só o facto da oferta estar em ciclo, bem como a sua obrigatoriedade e a propensão global do cliente à compra de produtos não obrigatórios, traduzindo-se numa melhoria na assertividade do produto oferecido ao cliente.

7. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Sendo este um projeto inovador na Fidelidade, houve desde cedo algumas limitações.

Numa primeira fase, e depois do desenho da solução NBA estar criado, percebeu-se que a companhia não recolhia toda a informação necessária para implementar um Sistema de Recomendação à altura do planeado. Posto isto, a primeira versão ficou desde logo comprometida.

Numa segunda fase, o facto de não haver uma catalogação das coberturas obrigatórias/ não obrigatórias e das respetivas regras que permitissem essa definição, levou a um impasse no processo de desenvolvimento do NBA, pois só foi possível concluir o modelo de propensão à compra depois da classificação de todas as coberturas dos produtos designados por lei como obrigatórios. Após o mapeamento de todas as coberturas e/ou produtos estar concluído, foi necessário transpor o mesmo na criação de tabelas que representassem esse mesmo mapeamento, para que a informação ficasse estanque e acessível para futuros reprocessamentos ou análises. Para isso, foi necessário realizar um levantamento exaustivo e moroso de todas as tabelas utilizadas pelas diversas áreas de negócio de modo a obter a informação da forma mais correta.

Em termos de recomendações futuras, e numa segunda versão do NBA, deverá revisitar-se o tema da definição de obrigatoriedade, tentando acrescentar mais detalhe ao mapeamento das coberturas para que o modelo de propensão à compra seja ainda mais preciso. Por exemplo, no ramo Automóvel, seria interessante verificar se um automóvel foi adquirido através de um empréstimo bancário, porque nesses casos, as coberturas ditas obrigatórias iriam aumentar. Outra situação a ter em conta será a idade do veículo pois, mesmo que o cliente seja sensível ao tema dos seguros, não conseguirá contratar todas as coberturas que deseja. No ramo das Embarcações será necessário analisar os clientes com capitais contratados de 350 000€ para a cobertura de Responsabilidade Civil, pois se o cliente navegar entre Portugal e Espanha, este capital é obrigatório por lei, por outro lado, um cliente que adquira esta cobertura com este capital mas navegue apenas em Portugal, traduz-se, à partida, num cliente mais preocupado. Para finalizar este tema, no ramo de Incêndio seria importante verificar se o imóvel foi adquirido através de um crédito habitação, e verificar o tipo de habitação, visto que a cobertura de incêndio apenas é obrigatória para a propriedade horizontal.

Em futuras versões será necessário criar robustez no sistema de recomendações, através da conclusão dos restantes pilares em falta: o da Retenção e das Ações e, por consequência, a criação do coeficiente de retenção em falta. Para além do desenvolvimento destes pilares, é essencial concluir o pilar dos Produtos & Serviços, adicionando as ofertas de *up-sell*. À medida que cada pilar é criado ou reforçado, existe a necessidade constante de perceber como é que as novas peças se irão encaixar no sistema, procedendo às mudanças necessárias na estrutura do mesmo e no cálculo do *score* final.

8. BIBLIOGRAFIA

- Aggarwal, V., & Kosian, S. (2011). *Feature Selection and Dimension Reduction Techniques in SAS®*. Obtido de <https://www.lexjansen.com/nesug/nesug11/sa/sa08.pdf>
- Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *IJCSI International Journal of Computer Science Issues*.
- Artigo 10º, Decreto-Lei n.º 315/2009. (s.d.). Obtido de Electrónico, Diário da República: https://dre.pt/web/guest/pesquisa/-/search/483402/details/normal?p_auth=VRzEbJt1&q=315%2F2009&y=5&x=9
- Artigo 12º, Decreto-Lei n.º 291/2007. (s.d.). Obtido de <https://dre.pt/pesquisa/-/search/640637/details/maximized>
- Artigo 1421º, Decreto-Lei n.º 267/94. (s.d.). Obtido de <https://dre.pt/application/dir/pdf1sdip/1994/10/247A00/64296433.pdf>
- Artigo 1429.º, Decreto-Lei n.º 267/94. (s.d.). Obtido de <https://dre.pt/application/dir/pdf1sdip/1994/10/247A00/64296433.pdf>
- Artigo 26º, Decreto-Lei n.º 235/92. (s.d.). Obtido de http://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=2944&tabela=leis&so_miolo=
- Artigo 4º, Decreto-Lei nº 291/2007. (s.d.). Obtido de <https://dre.pt/pesquisa/-/search/640637/details/maximized>
- Artigo 76º, Decreto-Lei n.º 24/2018. (s.d.). Obtido de http://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=2872&tabela=leis&ficha=1&pagina=1&so_miolo=
- Auto 1234. (s.d.). Obtido de <https://www.fidelidade.pt/PT/particulares/Auto/Produtos/auto-1234/Paginas/default.aspx>
- Auto Liber 3G. (s.d.). Obtido de https://www.fidelidade.pt/PT/particulares/Auto/Produtos/Auto/Documents/AU024_CG38_Seguro_Automovel_Liber_3G_nov2019.pdf
- Azevedo, C. S., & Santos, M. F. (2005). *Data Mining – Descoberta de Conhecimento em Bases de Dados*. FCA Publisher.
- ÇELİK, Ö., & ALTUNAYDIN, S. S. (2018). A Research on Machine Learning Methods and Its Applications. *Journal of Educational Technology & Online Learning*.
- D., T. B., Kirchner, A., Eck, A., & Signorino, C. S. (2018). An Introduction to Machine Learning Methods for Survey Researchers.
- Davenport, T. H., D'Almeida, L., & Lucker, J. (2011). Know What Your Customers Want Before They Do. *Harvard Business Review*.

- Decreto-Lei* n.º 84/2019. (s.d.). Obtido de http://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=3095&tabela=leis&ficha=1&pagina=1&so_miolo=
- deVille, B., & Padraic, N. (2013). *Decision Trees for Analytics Using SAS Enterprise Miner*. SAS Institute Inc.
- Família - Empregada Doméstica*. (s.d.). Obtido de https://www.fidelidade.pt/PT/particulares/Habitacao/Seguros/Documents/FM027_CG011_Acidente%20de%20Trabalho_Empregada%20Dome%CC%81stica_Jul2018.V1.pdf
- Fenjiro, Y. (s.d.). *Medium*. Obtido de <https://medium.com/@fenjiro/recommender-systems-d0e597424a98>
- Galetto, M. (2015). *ngdata*. Obtido de <https://www.ngdata.com/what-is-customer-retention/>
- Gilberto, F. (2012). *Manual Prático dos Seguros*. Lidel.
- Guido, J. J., Winters, P. C., & Rains, A. B. (2006). Logistic Regression Basics. *University of Rochester Medical Center, Rochester, NY*.
- Güneş, F., Tharrington, R., Abbey, R., & Hunt, X. (2020). How to Explain Your Black-Box Models in SAS Viya. *SAS Global Forum*.
- Gurbanov, T., & Ricci, F. (2017). Action Prediction Models for Recommender Systems Based on Collaborative Filtering and Sequence Mining Hybridization. *The 32nd ACM SIGAPP Symposium On Applied Computing*.
- Jornal Económico. (s.d.). *O desafio dos seguros na era digital*. Obtido de <https://jornaleconomico.sapo.pt/noticias/o-desafio-dos-seguros-na-era-digital-167710>
- Kamakura, W. A. (2007). Cross-Selling: Offering the Right Product to the Right Customer at the Right Time. *Journal of Relationship Marketing*.
- Lavery, R., & Mawr, B. (2016). *An Animated Guide: Deep Neural Networks in SAS® Enterprise Miner*.
- Linoff, G. S., & Berry, M. J. (2011). Chapter 7 - Decision Trees. Em *Data Mining Techniques For Marketing, Sales, and Customer*. Wiley Publishing, Inc.
- Linoff, G. S., & Berry, M. J. (2011). Chapter 8 - Artificial Neural Networks. Em *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley Publishing, Inc.
- Mitra, S., Chaudhari, N., & Patwardhan, B. (2014). Leveraging Hybrid Recommendation System in Insurance Domain. *International Journal of Engineering and Computer Science*.
- Molnar, C. (s.d.). *Interpretable Machine Learning*.
- Nichols, J. A., Chan, H. W., & Baker, M. A. (2018). Machine learning: applications of artificial intelligence to imaging and diagnosis.

Olaleye, D. (2019). Interpreting black-box machine learning models: Real-world evidence examples and applications. SAS.

Portaria n.º 689/2001. (s.d.). Obtido de <https://dre.pt/pesquisa/-/search/348547/details/maximized>

Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm. *arXiv*.

Rasouliyan, L., & Miller, D. P. (s.d.). *The Logic and Logistics of Logistic Regression Including New Features in SAS® 9.2*.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (s.d.). *Recommender Systems Handbook*. Springer.

Sarma, K. S., & Corp, E. R. (2007). *Variable Selection and Transformation of Variables in SAS Enterprise Miner 5.2*.

SAS. (s.d.). *Create a Gradient Boosting Model of the Data*. Obtido de <https://documentation.sas.com/?docsetId=emgsj&docsetTarget=p03iy98sk0c9bvn1r6x7p px8uj08.htm&docsetVersion=14.3&locale=en>

SAS. (s.d.). *Data Partition Node*. Obtido de <https://documentation.sas.com/?docsetId=emref&docsetTarget=n0u3s4tv5v88cfn1dx0zkf d9pjm5.htm&docsetVersion=14.3&locale=en>

SAS Institute. (s.d.). *Data Mining and the Case for Sampling*.

SAS. (s.d.). *Overview of Variable Clustering*. Obtido de <https://documentation.sas.com/?docsetId=vdmmlref&docsetTarget=n03yg60d2z6gobn1u nqbo4w4q7h2.htm&docsetVersion=8.2&locale=en>

SAS. (s.d.). *Random Forest Model*. Obtido de <https://documentation.sas.com/?docsetId=fcmrug&docsetTarget=n1tzufyosz12bhn12ls3i md63q94.htm&docsetVersion=14.2&locale=en>

SAS. (s.d.). *Regression Node*. Obtido de <https://documentation.sas.com/?docsetId=emref&docsetTarget=n1jqz8cssr9m2n1ktx2iy v87q56.htm&docsetVersion=14.3&locale=en#n09497077wxa3cn1o8v1eoihxwu5>

SAS. (s.d.). *The HPFOREST Procedure*. Obtido de https://documentation.sas.com/?docsetId=emhpprcref&docsetTarget=emhpprcref_hpfor est_details29.htm&docsetVersion=14.2&locale=pt

SAS. (s.d.). *The LOGISTIC Procedure*. Obtido de https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_logistic_details28 .htm&docsetVersion=15.1&locale=en

SAS. (s.d.). *Variable Selection Node*. Obtido de <https://documentation.sas.com/?docsetId=emref&docsetTarget=n1m7rvh6yyb3mmn0zav ezsher4ml.htm&docsetVersion=14.3&locale=en#p06glh3awamo21n1mn6ykd3k2m5n>

statistics, L. (s.d.). *Pearson Product-Moment Correlation*. Obtido de <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>

Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System. *International Journal of Computer Applications*.

Zhao, K. (s.d.). *Predictive Modeling Using Artificial Neural Networks in SAS® Enterprise Miner*.

9. ANEXOS

Anexo 1 - Glossário de Seguros

- Apólice: O conjunto de condições identificado na cláusula anterior e na qual é formalizado o contrato de seguro celebrado (Família - Empregada Doméstica, s.d.).
- Cabaz/Carteira: Conjunto de contratos de seguros ou dos contratos de capitalização subscritos junto de uma empresa de seguros (Gilberto, 2012).
- Capital seguro: Montante estipulado nas condições particulares do contrato como sendo o limite máximo de responsabilidade da empresa de seguros (Gilberto, 2012).
- Dano corporal: Prejuízo resultante de lesão da saúde física ou mental (Auto Liber 3G, s.d.).
- Dano material: Prejuízo resultante de lesão de coisa móvel, imóvel ou animal (Auto Liber 3G, s.d.).
- Mediador: Aquele que exerce a atividade de mediação de seguros (Gilberto, 2012).
- Pessoa Segura: O trabalhador por conta de outrem, ao serviço do Tomador do Seguro, titular do interesse seguro, bem como os administradores, diretores, gerentes ou equiparados, quando remunerados (Família - Empregada Doméstica, s.d.).
- Prémio Comercial: Custo teórico médio das coberturas do contrato, acrescido de outros custos, nomeadamente de aquisição e de administração do contrato, bem como de gestão e de cobrança (Gilberto, 2012).
- Ramo (de seguro) /LOB: Conjunto de operações ou atividades relativas a contratos de seguro da mesma natureza. Por exemplo: ramo incêndio, ramos mercadorias transportadas, ramo vida, etc. (Gilberto, 2012).
- Sinistro: A verificação, total ou parcial, do evento que desencadeia o acionamento das coberturas do risco previstas no contrato (Auto Liber 3G, s.d.).
- Tomador do Seguro: A entidade empregadora que contrata com o Segurador, sendo responsável pelo pagamento do prémio (Família - Empregada Doméstica, s.d.).

Anexo 2 – Valores dos impactos das variáveis selecionadas no modelo

		Pr > ChiSq	Odds	Effects	
Estado do cliente					
ANTIGO	RECUPERADO	<.0001	1,389	↑	38,9%
CONHECIDO	RECUPERADO		1,443	↑	44,3%
NOVO	RECUPERADO		0,423	↓	-57,7%
RECENTE	RECUPERADO		1,254	↑	25,4%
Segmento do cliente					
A1P - Human Seeker	S1P - Brand Oldie	<.0001	1,389	↑	38,9%
A2A - Lost in Translation	S1P - Brand Oldie		1,235	↑	23,5%
A2R - Online Price Grabber	S1P - Brand Oldie		0,991	↓	-0,9%
A3H - Wealthy Investor	S1P - Brand Oldie		1,297	↑	29,7%
J1R - Remote Low Budget	S1P - Brand Oldie		1,292	↑	29,2%
J2H - Rational	S1P - Brand Oldie		1,085	↑	8,5%
PE - Pequenas Empresas	S1P - Brand Oldie		2,432	↑	143,2%
PN - Pequenos Negócios	S1P - Brand Oldie		1,068	↑	6,8%
Cabaz atual					
AUTO	Outros	<.0001	1,107	↑	10,7%
AUTO / CAP	Outros		0,992	↓	-0,8%
AUTO / MRH	Outros		1,904	↑	90,4%
AUTO / MRH / PPR	Outros		1,196	↑	19,6%
AUTO / MRH / SD	Outros		1,663	↑	66,3%
AUTO / MRH / VDR	Outros		1,169	↑	16,9%
AUTO / PPR	Outros		1,184	↑	18,4%
AUTO / SD	Outros		1,120	↑	12,0%
AUTO / VDR	Outros		1,196	↑	19,6%
MRH	Outros		1,662	↑	66,2%
Cabaz há 12 meses					
AUTO	SD	<.0004	0,911	↓	-8,9%
AUTO / CAP	SD		1,159	↑	15,9%
AUTO / MRH	SD		0,709	↓	-29,1%
AUTO / MRH / VDR	SD		0,982	↓	-1,8%
AUTO / PPR	SD		0,744	↓	-25,6%
AUTO / SD	SD		0,999	↓	-0,1%
MRH	SD		0,728	↓	-27,2%
MRH / VDR	SD		1,234	↑	23,4%
Outros	SD		1,011	↑	1,1%
PPR	SD		0,88	↓	-12,0%

		Pr > ChiSq	Odds	Effects	
Família da última anulação efetiva da apólice					
AP	VDR	<.0001	0,936	↓	-6,4%
ATCO	VDR		0,684	↓	-31,6%
ATCP	VDR		0,536	↓	-46,4%
AUTO	VDR		0,648	↓	-35,2%
CACA	VDR		0,604	↓	-39,6%
CAP	VDR		0,692	↓	-30,8%
EMB	VDR		0,484	↓	-51,6%
MRH	VDR		0,829	↓	-17,1%
MRN	VDR		0,449	↓	-55,1%
NVDO	VDR		1,054	↑	5,4%
PPR	VDR		0,766	↓	-23,4%
RCF	VDR		0,565	↓	-43,5%
RCG	VDR		0,634	↓	-36,6%
SD	VDR		0,762	↓	-23,8%
VDO	VDR		1,487	↑	48,7%
Família de produtos da última apólice comprada					
AP / AUTO	RCG	<.0001	0,055	↓	-94,5%
ATCO	RCG		0,001	↓	-99,9%
AUTO	RCG		0,06	↓	-94,0%
AUTO / CAP	RCG		0,043	↓	-95,7%
AUTO / MRH	RCG		0,214	↓	-78,6%
AUTO / SD	RCG		0,154	↓	-84,6%
AUTO / VDR	RCG		0,050	↓	-95,0%
CACA	RCG		0,021	↓	-97,9%
MRH	RCG		0,108	↓	-89,2%
Outros	RCG		0,081	↓	-91,9%
RCF	RCG		45,271	↑	4427,1%
Família da apólice em vigor mais recente					
ATCO	SD	<.0001	0,403	↓	-59,7%
AUTO	SD		1,866	↑	86,6%
AUTO / CAP	SD		6,113	↑	511,3%
AUTO / MRH	SD		2,352	↑	135,2%
AUTO / VDR	SD		1,971	↑	97,1%
CACA	SD		0,378	↓	-62,2%
CAP	SD		8,637	↑	763,7%
MRH	SD		1,671	↑	67,1%
Outros	SD		2,149	↑	114,9%
RCF	SD		11,684	↑	1068,4%
RCG	SD		459,626	↑	45862,6%
Fracionamento de pagamento preferencial					
Anual	Único	<.0001	1,199	↑	19,9%
Mensal	Único		1,202	↑	20,2%
Semestral	Único		1,413	↑	41,3%
Trimestral	Único		0,900	↓	-10,0%

		Pr > ChiSq	Odds	Effects	
Descrição do Mediador / Agente Preferencial					
CGD	OUTROS	<.0001	0,291	↓	-70,9%
DIR COMERCIAL AGÊNCIAS	OUTROS		0,814	↓	-18,6%
DIR COMERCIAL I	OUTROS		1,058	↑	5,8%
DIREÇÃO MEDIAÇÃO NORTE	OUTROS		0,991	↓	-0,9%
DIREÇÃO MEDIAÇÃO SUL	OUTROS		1,033	↑	3,3%
Idade					
		<.0001	1,016	↑	1,6%
Porcentagem de apólices em vigor com fracionamento único					
		0.0421	1,516	↑	51,6%
Qtd. Apólices efetivamente anuladas nos últimos 12 meses no ramo Automóvel					
		<.0001	0,802	↓	-19,8%
Qtd. Apólices novas criadas nos últimos 12 meses no ramo Automóvel					
		<.0001	0,127	↓	-87,3%
Qtd. Apólices novas criadas nos últimos 12 meses no ramo Incêndio					
		<.0001	0,025	↓	-97,5%
Qtd. Apólices em vigor há mais de 8 anos					
		0.0009	1,036	↑	3,6%
Qtd Apólices em vigor onde o cliente é tomador Automóvel					
		<.0001	0,894	↓	-10,6%
Qtd Apólices em vigor onde o cliente é tomador Responsabilidade Civil Familiar					
		0.0057	1,348	↑	34,8%
Saldo de Apólices nos últimos 12 meses no ramo Multi-Riscos Habitação					
		0.0002	0,853	↓	-14,7%
Sexo					
F vs M		<.0001	1,253	↑	25,3%
Valor do Prêmio Comercial Anualizado das apólices novas criadas nos últimos 12 meses no ramo Automóvel					
		<.0001	1,012	↑	1,2%

